A STUDY OF THE TAIL MEASURE AND ITS APPLICATIONS IN RISK MODELING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Duc Hoai Quyen Nguyen

August 2013

© 2013 Duc Hoai Quyen Nguyen
ALL RIGHTS RESERVED

A STUDY OF THE TAIL MEASURE AND ITS APPLICATIONS IN RISK MODELING

Duc Hoai Quyen Nguyen, Ph.D.

Cornell University 2013

This dissertation has 4 chapters, in which we attempt to explore and analyze the structure of extremal data.

The first chapter is a review of statistical estimation methods of the tail in the context of extreme value theory as well as their applications in risk managmenet.

The quality of estimation of multivariate tails depends significantly on the portion of the sample included in the estimation. Hence, the second chapter describes an approach involving sequential statistical testing in order to select which observations should be used for estimation of the tail. The method is computationally efficient, and can be easily automated. No visual inspection of the data is required. The consistency of the Hill estimator is established when used in conjunction with the proposed method, as well as its asymptotic fluctuations.

The third chapter expands the previous method to the multivariate case. The estimator for the tail measure is proven to be consistent using this method of tail selection. We test the proposed method on simulated data, and subsequently apply it to analyze CoVaR for stock and index returns.

Finally, we study the structure of spectral measures in financial data. We make observations about certain characteristics of the measures and subsequently propose an approach that can help us study the spectral measure in the face of high dimensional sparsity.

BIOGRAPHICAL SKETCH

Duc Nguyen, also known as Tilo Nguyen, was born and grew up in Saigon, Vietnam. She graduated Magna cum Laude from Smith College in May 2007 with a B.A. in Mathematics. Later that year, she joined the Center for Applied Mathematics at Cornell University as a Ph.D. student, with concentration in Probability and Statistics.

This document is dedicated to my family.

Your constant love and support

made this possible.

ACKNOWLEDGEMENTS

I am tremendously grateful to my advisor, Gennady Samorodnitsky, who guided me every step of the way through my research and is vital to my success at Cornell. His patience, unwavering support and invaluable inputs has brought my dissertation to fruition. I'm also grateful to Professor Laurent Saloff-Coste and Professor Ping Li for being my committee members.

My research was partially supported by the Army Research Office (ARO) grants W911NF-07-1-0078 and W911NF-12-10385, National Science Foundation (NSF) grants DMS-1005903, DMS-0739164 and National Security Agency (NSA) grant H98230–11-1-0154 at Cornell University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

EXTREME VALUE THEORY AND RISK MANAGEMENT

1.1 A general introduction to extreme financial data and extreme value theory

In 1963, Mandelbrot made a major contribution by demonstrating that the normal distribution was a poor fit for certain financial returns ([Mandelbrot, 1963]). His findings on asset returns were followed by work by Eugene Fama and Richard Roll and resulted in the formation of the Stable Paretian model. Today, it is accepted as fact that not only do financial returns possess a much heavier tail than the normal distribution, they also have a higher degree of "peakedness" at the center and are asymmetric. Other stylized facts of asset returns include volatility clustering and dependence among data.

Distributional assumptions about asset returns are important as they play a key role in critical tasks such as pricing options, risk forecasting and portfolio optimization.

From a risk management perspective, one needs models that adequately predict the chance and consequences of rare events, which by definition does not lend to a large amount of data that modelers can use. For example, to quantify market risks, institutions use popular measures such as Value at Risk or Expected Shortfall to quantify the extreme losses that can possibly occur due to market movements. These measures are usually based on the 5% or 1% quantile of the return distribution, which is typically near the edge of the range of available data. For credit or operational risks, we need to model extreme losses from defaults, credit rating

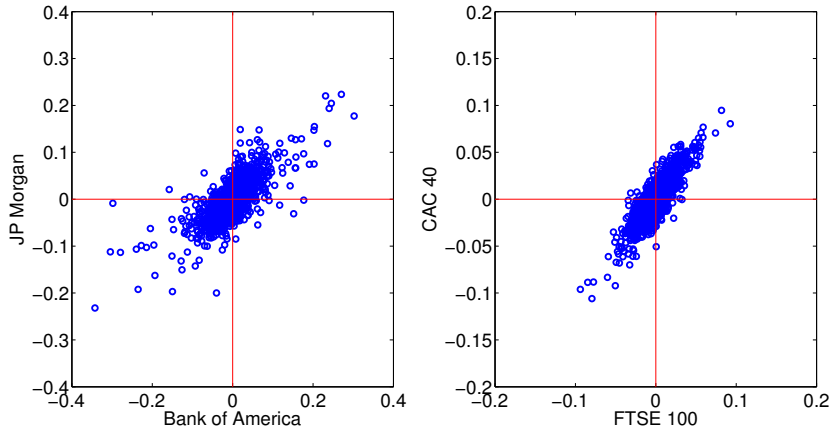

Figure 1.1: Daily log return

downgrades or potential operational problems. To adequately perform these tasks, one needs to make certain assumptions about the loss distribution to be able to extrapolate beyond the existing data.

Beyond the univariate setting, it is important to note that certain financial instruments, related by industry, or by geography, tend to show highly dependent extreme movements. Figure 1.1 shows a scatter plot of the daily log returns on the stock of Bank of America and JP Morgan over a period of 4362 days from January 1996 to June 2012, next to a scatter plot of daily log returns on the British FTSE 100 index and the French CAC 40 index, from January 1995 to April 2012.

It is obvious to the "naked eye" that the two indices have higher dependence in the extreme movements than the two stocks do. The difficulty with making this precise via statistical estimation is two-fold. First, there are simply not that many extreme observations. Second, it is not easy to decide in a given sample what observations "belong to the tail region" and should be used to estimate the dependence of the extremes.

Why such an understanding of multivariate extremal dependence matters to

risk management is obvious. But another area of finance where it also plays a major role is the problem of portfolio choice. The reason stems from one of the most accepted principles of finance: high risks, high returns. Investment decisions have to be made with both risks and returns fully taken into account. This is thus the basis of portfolio choice theory. A portfolio choice problem is typically a problem of maximizing a utility function subjected to some constraints such as risk measures. One of the first rigorous works on portfolio selection problem is by Harry Markowitz ([Markowitz, 1952], [Markowitz, 1959]). He described the construction of an optimal portfolio with reward of a portfolio as the expected return and used the variance of returns as a measure of risks. His work was then extended to an equilibrium theory and the famous CAPM model by [Sharpe, 1964], [Lintner, 1965] and [Mossin, 1966]. Even in his pioneering work, the shortcomings of variance as a risk measure was pointed out by Markowitz. Note that this was before literature started discarding the normal distribution as a viable assumption, thus he proposed to replace it with a quantity called semi-variance, which emphasized only the losses rather than movements in both directions.

It is understandable that as researchers moved away from the Gaussian models to better capture the heavy tails of univariate financial data, the same effort would be extended to multivariate risks. The assumption that several assets may suffer strong losses at the same time is not unreasonable. [Barro, 2006] studied economic disasters in the twentieth century and estimates that the probability for a disaster in any year is between 1.5 and 2 percent. He defined an event as an economic disaster if the real GDP per capita suffered a contraction by at least 15 percent. This belief that extreme comovements of asset returns can not be neglected is also played out on stock exchanges. Data on March 28, 2013 showed that it costs \$85 per option contract to protect against a 10% drop in the S&P 500 through

mid-April.

Extreme value theory arose as a useful tool to model extreme financial data in both univariate and multivariate settings as described above. It contains a broad class of distributions which resemble the stylized facts of financial returns as well as give a framework to model dependence in extreme movements. Extreme value theory can generally be broken into two areas: analyzing the maximum of a set of data (see [Embrechts et al., 1997] for an extensive review), and a topic of focus in this thesis - the study of extreme data past a chosen threshold. There are two different ways to model extreme data in this second context: using the concept of regular variation (a semi parametric approach) or a completely parametric one.

Roughly speaking, regular variation means the distribution function has tails that fall off like a power law. Its definition will be stated formally in the next section. There are several advantages that comes with using the theory of regular variation. First, a large class of popular heavy tailed distributions, which includes Pareto, Student's t, stable distribution as well as their multivariate versions, satisfies this description of power-law like behavior. The second advantage is that regular variation assumption only focuses on the tail behavior which is perfect for risk models. In contrast, parametric modeling of the whole distribution requires fitting the non-extreme data, which in turn may results in a suboptimal fit for the tail. Lastly, regular variation leads to a self similarity structure in the tail that we can use to extrapolate beyond existing data.

1.2 Semiparametric methods for modeling extreme data

1.2.1 Regular variation

Let F be a univariate distribution function. F is said to have a regularly varying right tail of index $\alpha > 0$ if the tail probability function $\bar{F} = 1 - F$ satisfies

$$\lim_{x\to\infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = t^{-\alpha} \tag{1.1}$$

for all $t > 0$. The index α measures the heaviness of the tail. The smaller is α, the heavier is the right tail of the distribution. An encyclopedic treatment of regular variation is given by [Resnick, 1987, 2007].

The concept of regular variation can be extended to the multivariate case. Let $\mathbf{Z}$ be a d-dimensional random vector with nonnegative components. Denote $\mathbb{E} = [0,\infty]^d \setminus \{\mathbf{0}\}$. We say that $\mathbf{Z}$ is multivariate regularly varying if there exist a function $b(t) \to \infty$ and a nontrivial (i.e. nonzero) Radon measure ν on $\mathbb{E}$, vanishing on the set of infinite points, such that

$$tP\left[\frac{\mathbf{Z}}{b(t)} \in \cdot\right] \xrightarrow{v} \nu, \quad t \to \infty\,,$$

vaguely in the space of nonnegative Radon measures on $\mathbb{E}$; see [Resnick, 2007]. The measure ν is usually called the tail measure. The index α of the regular variation of the tail in this definition is hidden in the index of regular variation of the function b. It also appears in the scaling property the limit measure ν (the so-called tail measure) must possess: $\nu(c\cdot) = c^{-\alpha}\nu(\cdot)$ for every $c > 0$.

Equivalently, $\mathbf{Z}$ is multivariate regularly varying with index $\alpha > 0$ if there exists a probability measure S on $\mathbb{S}_+ = \mathbb{S}^{d-1} \cap [0,\infty)^d$ such that for all $x > 0$

$$t\mathbb{P}\Big(||\mathbf{Z}|| > b(t)x, \frac{\mathbf{Z}}{||\mathbf{Z}||} \in \cdot\Big) \Rightarrow cx^{-\alpha}S(\cdot), \quad t \to \infty\,,$$

weakly on the unit sphere $\mathbb{S}^{d-1}$; see [Resnick, 2007]. The probability measure S is called the spectral measure of the distribution. One can choose an arbitrary norm on $\mathbb{R}^d$; this will determine the unit sphere and, thus, the spectral measure. The constant c depends on the choice of the normalizing function b.

Clearly, if the same normalizing function b is used in both definitions, then the tail and spectral measures are related by

$$\nu\Big(||\mathbf{Z}|| > x, \frac{\mathbf{Z}}{||\mathbf{Z}||} \in \cdot\Big) = cx^{-\alpha}S(\cdot),\ x > 0\,. \tag{1.2}$$

If the tail measure (or the spectral measure) is concentrated on the axes, it is common to speak of asymptotic (or tail) independence between the components of the random vector $\mathbf{Z}$, in the sense of very low likelihood that more than one component takes a large value at the same time. Otherwise, the components of the vector $\mathbf{Z}$ are said to be tail dependent. A special example of tail independence is the case of a random vector with i.i.d. regularly varying marginal distributions.

We should note that these definitions are really useful if all the marginal tails are equivalent, which means that they are regularly varying with the same index α. In practice this is often an unreasonable assumption. There exists a framework of multivariate regular variation that allows different marginal tail indices. Let $\mathbf{Z}$ still be a d-dimensional random vector in the nonnegative orthant. Specifically, we assume that there exist d normalizing functions $b^{(i)}(t) \to \infty$, $i = 1 \ldots, d$, such that both

$$t\,\mathbb{P}\left[\left(\frac{Z^{(i)}}{b^{(i)}(t)}\right) \in \cdot\right] \xrightarrow{v} \nu_{\alpha_i}, \quad t \to \infty \tag{1.3}$$

vaguely on $[0, \infty)$ for each $i = 1, \ldots, d$ and

$$t\,\mathbb{P}\left[\left(\frac{Z^{(1)}}{b^{(1)}(t)}, \frac{Z^{(2)}}{b^{(2)}(t)}, \ldots, \frac{Z^{(d)}}{b^{(d)}(t)}\right) \in \cdot\right] \xrightarrow{v} \nu, \quad t \to \infty \tag{1.4}$$

vaguely on $\mathbb{E}$, where ν_{α_i} is a measure on the Borel sets $[0,\infty)$ with the density $c_i\alpha_i x^{-(1+\alpha_i)}$, $x>0$, α_i, $c_i>0$, $i=1,\dots,d$ with respect to the Lebesgue measure. Moreover, ν is a nontrivial Radon measure on $\mathbb{E}$, vanishing on the set of infinite points, i.e. points which have ∞ in at least one coordinate; it is also called the tail measure. The ith normalizing function b_i is regularly varying with index $1/\alpha_i$, $i=1,\dots,d$. Since the indices of regular variation may be different in different directions, the normalizing functions may be different as well; this is the nonstandard version of regular variation. It can be converted into the standard version of regular variation, with equivalent tails, as follows. Let F_i, $i=1,\dots,d$ be the marginal distribution functions. Define

$$u_i(x)=\left(\frac{1}{1-F_i}\right)^{\leftarrow}(x),\ i=1,\dots,d\,. \tag{1.5}$$

$F^{\leftarrow}$ stands for the inverse function of F. Then

$$t\,\mathbb{P}\left[\left(\frac{u_i^{\leftarrow}\big(Z^{(i)}\big)}{t},\ i=1,\dots,d\right)\in\cdot\right]\xrightarrow{v}\nu_*(\cdot) \tag{1.6}$$

vaguely on $\mathbb{E}$, while for each $i=1,\dots,d$,

$$t\,\mathbb{P}\left[\frac{u_i^{\leftarrow}\big(Z^{(i)}\big)}{t}>x\right]\to x^{-1},\ x>0\,. \tag{1.7}$$

Here ν_* is a nontrivial Radon measure on $\mathbb{E}$, vanishing on the set of infinite points, with the scaling property

$$\nu_*(c\cdot)=c^{-1}\nu_*(\cdot),\ c>0\,. \tag{1.8}$$

The measures ν and ν_* are related by

$$\nu([\mathbf{0},\mathbf{x}^{\mathbf{1}/\alpha}]^c)=\nu_*([\mathbf{0},\mathbf{x}]^c)\,,$$

where $\mathbf{x}^{\mathbf{1}/\alpha}$ is defined as the vector with coordinates x_i^{1/α_i}, $i=1,\dots,d$.

Therefore, the transformation (1.5) achieves both the standard global regular variation and, asymptotically, the standard Pareto index 1 marginal tails (if F_i is

continuous, then, after the transformation, the ith marginal is, in fact, the standard index 1 Pareto distribution).

1.2.2 Statistical estimations of tail index and tail measure

Estimators of the tail index

Estimating the tail index α is of crucial importance in many applications of stochastic models, and a number of estimators have been designed for that purpose. We will give here an overview of the most popular estimators.

The best known estimator of the tail index is the Hill estimator, introduced by [Hill, 1975], and it is defined as follows. Assume that $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ are the order statistics from a positive sample (or from the positive part of a general sample) $X_1, \ldots, X_n$. The Hill estimator based on k upper order statistics is defined as

$$H_{k,n} := \frac{1}{k} \sum_{i=0}^{k-1} \log \frac{X_{n-i,n}}{X_{n-k,n}} \tag{1.9}$$

If the sample is an i.i.d. sample from a distribution with a regularly varying right tail with tail index α, then, under the conditions $n \to \infty$, $k \to \infty$ and $\frac{k}{n} \to 0$, the Hill estimator $H_{k,n}$ converges in probability to $\gamma = \frac{1}{\alpha}$ (see [Mason, 1982]). If, additionally, $k/\log\log n \to \infty$, then even almost sure convergence holds ([Deheuvels et al., 1988]). The role of the condition $\frac{k}{n} \to 0$ is to ensure that only data from the tail enter into the estimation. It has also been shown that the Hill estimator remains consistent under certain deviations from the i.i.d. assumption; a summary is in Theorem 6.4.6 of [Embrechts et al., 1997].

Another estimator for the tail index is the moment estimator, see [Dekkers and

de Haan, 1989; Dekkers et al., 1989; de Haan and Ferreira, 2006] . Suppose

$$M_{k,n}^{(i)} := \frac{1}{k}\sum_{i=0}^{k-1}\left(\log\frac{X_{n-i,n}}{X_{n-k,n}}\right)^i$$

Then the moment estimator is defined as

$$\hat{\gamma} = M_{k,n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_{k,n}^{(1)})^2}{M_{k,n}^{(2)}}\right)^{-1}$$

Assume the sample is i.i.d and comes from a distribution in the domain of attraction of an extreme value distribution,

$$G_\gamma(x) = \exp(-(1+\gamma x)^{-\frac{1}{\gamma}})$$

In the case of $\gamma > 0$, this assumption is equivalent to the assumption of regular variation, see [de Haan and Ferreira, 2006]. Under the conditions $n \to \infty$, $k \to \infty$ and $\frac{k}{n} \to 0$, the moment estimator $H_{k,n}$ converges in probability to γ (see [Dekkers et al., 1989]). If, additionally, $k/(\log n)^\delta \to \infty$ for some $\delta > 0$, then even almost sure convergence holds ([Dekkers et al., 1989])

A different estimator also defined within the framework of domain of attraction of G_γ is the Pickands estimator, see [Pickands, 1975]. It is defined as followed

$$\hat{\gamma} = \frac{1}{\log 2}\log\left(\frac{X_{n-k,n} - X_{n-2k,n}}{X_{n-2k,n} - X_{n-4k,n}}\right)$$

Again, weak consistency is established if we have $n \to \infty$, $k \to \infty$ and $\frac{k}{n} \to 0$. Strong consistency is established under the extra condition of $k/\log\log n \to \infty$, same as the Hill estimator.

Asymptotic convergence to the normal distribution for all of the above estimators are established under extra conditions on k and the underlying distribution, in particular a second order regular variation condition, see [Embrechts et al., 1997; de Haan and Ferreira, 2006]. In practice, the second order regular variation is not easy to verify.

Comparing their asymptotic behavior, no one estimator performs superiorly across all distributions [de Haan and Peng, 1998] . Depending on the specific index γ and an index of second order regular variation of the underlying distribution (which we will define later), one out of these three estimators would give us a smaller asymptotic mean square error than the other two. Neveretheless, one advantage that the Pickands estimator has compared to the other two is that it is invariant with respect to changes in scale and location. However, all results on asymptotic behavior of tail index estimators were based on the assumption of i.i.d. samples which is unrealistic when it comes to financial data. [Kearns and Pagan, 1997] studied several sets of dependent time series data and concluded that asymptotic theory severely underestimated the actual standard errors associated with the estimators. From their studies, by comparing biases and variances, they concluded that out of the three estimators, the Hill estimator seemed to be the best choice.

Note that the moment estimator and the Pickands estimator are defined for all values of γ, while the Hill estimator is only defined in the case of heavy tail i.e $\gamma > 0$. For this reason, a Pickands or moment estimate of $\gamma \leq 0$ is an indication that a heavy tailed model may not be appropriate for the fitted data. Hence Pickands estimator and moment estimator are sometimes considered to be more informative in detecting if a heavy tail model is a good match. [Resnick, 2007] provided a few examples of this phenomenon.

Spectral measure estimations

Estimating the tail measure ν (or ν_*) is of crucial importance in many applications of stochastic models, and a number of estimators have been designed for that

purpose. A nonparametric method was proposed in [Einmahl et al., 2001] that automatically standardizes the problem without requiring one to estimate first the marginal quantile functions u in (1.5). This method also bypasses the need for estimation of each of the marginal tail indices.

Suppose that conditions (1.3) and (1.4) for global and marginal regular variation hold. Given a d-dimensional sample of size n, for each $j = 1, \ldots, d$, we define

$$r_i^j = \sum_{m=1}^{n} 1_{[Z_m^{(j)} \geq Z_i^{(j)}]}, \quad 1 \leq i \leq n\,,$$

which is the sequence of ranks. If $k = k(n)$ is an intermediate sequence, i.e. if $k \to \infty$ and $\frac{k}{n} \to 0$ as $n \to \infty$ then weak consistency of the empirical estimator

$$\frac{1}{k} \sum_{i=1}^{n} \epsilon_{\left(\frac{k}{r_i^{(j)}}, j=1,\ldots,d\right)} \Rightarrow \nu_* \tag{1.10}$$

in the vague topology on the space of nonnegative Radon measures on $\mathbb{E}$ holds; see pp. 311-312 in [Resnick, 2007]. Note that the condition $\frac{k}{n} \to 0$ ensures that only "tail observations" affect the estimator. The second order behavior of this estimator is difficult; in the two-dimensional case, and under additional regularity assumptions, asymptotic normality was established in [Einmahl et al., 2001].

To estimate the spectral measure, we first apply the polar transformation on the rank data, by

$$(R_{i,k}, \theta_{i,k}) = \left(\left\| \left(\frac{k}{r_i^{(j)}}, j = 1, \ldots, d \right) \right\|, \frac{\left(\frac{k}{r_i^{(j)}}, j = 1, \ldots, d \right)}{\left\| \left(\frac{k}{r_i^{(j)}}, j = 1, \ldots, d \right) \right\|} \right), \; i = 1, \ldots, n\,. \tag{1.11}$$

The continuity of the polar transformation and (1.2) and (1.10) imply

$$\frac{1}{k} \sum_{i=1}^{n} 1_{\left(R_{i,k} > 1,\, \theta_{i,k} \in (\cdot)\right)} \Rightarrow c S_*(\cdot) \tag{1.12}$$

weakly on the unit sphere, where S_* is the spectral measure corresponding to ν_* and c is a normalizing constant so that S_* is a probability measure. This then gives us the following consistent estimator for S_*:

$$S_{*k,n}(\cdot) := \frac{\sum_{i=1}^{n} 1_{\left(R_{i,k}>1,\, \theta_{i,k}\in(\cdot)\right)}}{\sum_{i=1}^{n} 1_{\left(R_{i,k}>1\right)}} \Rightarrow S_*(\cdot)\,. \tag{1.13}$$

For asymptotic normality of this estimator, see [Einmahl et al., 2001].

A modified version for the spectral measure estimator above was proposed by [Einmahl and Segers, 2009]. In bivariate context, they used the following marginal distributions of ν_* and the relationships between ν_* and S_*

$$\nu_*([0,\infty]\times[z,\infty]) = \nu_*([z,\infty]\times[0,\infty]) = \frac{1}{z}, 0 < z \leq \infty$$

$$\int_{\mathbb{E}} f\, d\nu_* = c\int_0^{\pi/2}\int_0^{\infty} f(z_1,z_2) r^{-2}\, dr\, S_*(d\theta)$$

where $f : \mathbb{E} \to \mathbb{R}$ is any ν_* integrable function, $z_1 = \dfrac{r\sin(\theta)}{||\sin(\theta),\cos(\theta)||}$, $z_2 = \dfrac{r\cos(\theta)}{||\sin(\theta),\cos(\theta)||}$. These gives rise to the following moment constraint on the spectral measure S_*

$$\int_0^{\pi/2} \frac{\sin(\theta)}{||\sin(\theta),\cos(\theta)||}\, S_*(d\theta) = \int_0^{\pi/2} \frac{\cos(\theta)}{||\sin(\theta),\cos(\theta)||}\, S_*(d\theta)$$

$$\Leftrightarrow \int_0^{\pi/2} \frac{\sin(\theta)-\cos(\theta)}{||\sin(\theta),\cos(\theta)||}\, S_*(d\theta) = 0$$

Let $g(\theta) = \dfrac{\sin(\theta)-\cos(\theta)}{||\sin(\theta),\cos(\theta)||}$ and $I_n = \{i = 1,...,n : R_{i,k} > 1\}$. The spectral

measure estimator is then the solution to the following optimization problem

$$\begin{aligned} \text{maximize } & \prod_{i \in I_n} p_i \\ \text{constraints } & p_i \geq 0 \\ & \sum_{i \in I_n} p_i = 1 \\ & \sum_{i \in I_n} g(\theta_{i,k}) p_i = 0 \end{aligned}$$

1.2.3 Applications in risk management

Recall that for $0 < q < 1$, and close to 1, $VaR_q^{(j)}$ is defined as the q-quantile of the loss distribution $F^{(j)}$ of the jth asset. To be precise,

$$VaR_q^{(j)} = F^{(j)\leftarrow}(q)$$

Suppose the loss distribution follows a regular variation assumption. Assume that $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ are the order statistics from the losses. Then as n gets very large, given a high threshold $X_{k,n}$, we have the following rough estimation for $x > X_{k,n}$

$$\frac{\bar{F}(x)}{\bar{F}(X_{k,n})} \approx \left(\frac{x}{X_{k,n}}\right)^{-\alpha} \tag{1.14}$$

We can then replace $\bar{F}(X_{k,n})$ by the empirical estimation $\dfrac{k-1}{n}$. Then an estimation for VaR_q when $q > 1 - \dfrac{k}{n}$ is as followed

$$\hat{VaR}_q(X) = X_{k,n} \left(\frac{n}{k-1}(1-q)\right)^{1/\hat{\alpha}}$$

This estimation of VaR depends on the choice of $X_{k,n}$ and the estimated $\hat{\alpha}$.

The estimation of the tail index α is also a routine task carried out daily by many risk departments for various financial products, partly to enable early detection of any abnormal pattern in market movements. There is significant interest in the literature on the tail index itself. Depending on different methods and different data sets used, there is a general debate on how heavy the tails of financial data are, ranging from an estimated α of 1 to 5. This debate may have major implications in the kind of models employed since it determines whether financial data have finite higher order moments or not. [Mandelbrot, 1963] studied the price process of cotton and suggested an α of 1.7. [Jansen and De Vries, 1991] used the limit laws of maxima to study certain stocks and indices return and concluded with α between 3 and 5. Their study indicated that financial returns would be better fitted with a t-distribution or GARCH model instead of the stable distribution. [Rachev and Mittnik, 2000] use Stable models to come up with α between 1 and 2 for stocks, indices and exchange rates. Other work on this topic includes [Loretan and Phillips, 1994], [McCulloch, 1996] and [Rachev et al., 2005].

[Danielsson and de Vries, 1997] imposed a second order condition on regular variation to get the following expansion $F(x) \approx 1 - ax^{-\alpha}[1 + bx^{-\beta}]$. They used an estimator for β proposed in [Danielsson and Vries, 1996] and the Hill estimator to calculate extreme quantiles for 2 foreign exchange rates. Analysis was done on both second returns and 10-minute returns over a period of one year. Later, [Danielsson and De Vries, 2000] combined this method with historical simulations to study VaR of portfolios of stocks and options. They came to the conclusion that at the 5% level, the famous RiskMetrics model did a better job but for a worse level of outcomes, RiskMetrics underestimated the VaR compared to their extreme value theory approach.

Regulators usually require banks to hold adequate capital to cover losses over a 10 day holding period in 99% of the occasions. Whereas, for internal risk management, banks typically use a holding period of 1 day and a confidence level of 5%. Hence we want to be able to calculate VaR estimate for returns over a period of several days from one-day VaR. For the calculation of this multi-time-unit VaR, a Gaussian based model uses the additivity of the normal distribution and implies $VaR_q(T) = T^{\cdot 5} VaR_q(1)$. Here $VaR_q(t)$ stands for the Value at Risk over a t-unit time period. This scaling property is no longer correct in a heavy tailed world. [Feller, 1971] showed that if $P(X > x) \approx ax^{-\alpha}$ for x large then

$$P(X_1 + X_2 + ... + X_T > x) \approx Tax^{-\alpha}$$

This implies a scaling factor of $T^{1/\alpha}$ for a T-period VaR.

Being able to estimate extreme quantiles well plays an important role in portfolio selection methods based on downside risk criteria. Roy's safety first criterion ([Roy, 1952]) is a portfolio choice problem where the criterion is that the probability of the returns falling below a desired benchmark is minimized. [Arzac and Bawa, 1977] later modified the problem of safety first investor into which one maximizes the expected return subjected to a constraint that the probability of failure past a prespecified threshold is no greater than a fixed number. [Jansen et al., 2000] used Equation 1.14 to calculate failure probabilities and then proceeded to choose the optimal portfolio among many hypothetical ones based on the above criterion. [Susmel, 2001] worked on a similar problem but with applications to the Latin American equity markets.

Even though the use of extreme value theory in portfolio choice significantly improves the estimate of failure probabilities, it has its own shortcoming. Note that if there are two assets, one has a lower tail index α than the other, any

convolution of the two assets would end up having the lower α as its tail index (see [Geluk and de Haan, 1987]). This effect, in turn, causes the safety first investor problem to end up in some cases with an unbalanced solution in which the asset with the highest tail index is favored above any other combinations. [Hyung and de Vries, 2007] then proposed a second order condition that would help mitigating the aforementioned trivial solutions .

Regular variation has also been used to model dependency and portfolio diversification. Studies on diversification and its effect on portfolio VaR in the presence of heavy tailed returns were done by [Hyung and de Vries, 2005] and [Ibragimov and Walden, 2007]. Multivariate regular variation was recently used to analyze extremal dependency among exchange rates, see [Hauksson et al., 2001] and [Stărică, 1999].

1.3 Parametric methods for modeling extreme data

1.3.1 The generalized Pareto distribution

The cumulative distribution function of the generalized Pareto distribution is given by

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi(y-\mu)}{\sigma}\right)^{-1/\xi} & ,\xi \neq 0 \\ 1 - \exp(-\dfrac{y-\mu}{\sigma}) & ,\xi = 0. \end{cases}$$

for $y > \mu$ when $\xi \geq 0$ and $\mu \leq y \leq \mu - \dfrac{\sigma}{\xi}$ when $\xi < 0$. This distribution is regularly varying with tail index $\alpha = \dfrac{1}{\xi}$ when $\xi > 0$ hence this case is our main interest from a risk modeling perspective.

The GPD can be fitted by using the maximum likelihood method, the method of moments or the method of probability weighted moments. For a comparison of these methods, see [Hosking and Wallis, 1987] and [Rootzén and Tajvidi, 1997].

The generalized Pareto distribution (GPD) is important in risk estimation because it serves as a natural model for excess probability over a high threshold. This stems from the following result. Let $\bar{F}_u(x) = P(X - u > x | X > u)$. [Pickands, 1975] proved that for every $\xi \in \mathbb{R}$, $F \in MDA(H_\xi)$ if and only if there exists $\sigma(u) > 0$ such that $\lim_{u \to x_F} \sup_{0<x<x_F-u} |F_u(x) - G_{\xi,\sigma(u)}(x)| = 0$. Here $MDA(H_\xi)$ stands for the maximum domain of attraction of an extreme value distribution with scale parameter ξ and x_F is the right end point of the distribution F. For more background on this topic, see [Embrechts et al., 1997].

The main point of the theorem is that for a large class of distributions, the excess distribution function F_u over a sufficiently high threshold u can be approximated by a GPD.

1.3.2 Applications in risk management

The generalized Pareto distribution and Pickands' theorem stated above provide a very simple estimation of the quantile function. From the definition of F_u we can write

$$\bar{F}(x) = \bar{F}(u)\bar{F}_u(x - u)$$

We can then approximate $\bar{F}$ with the empirical distribution function and $\bar{F}_u$ with GPD. This gives us the following quantile estimate

$$\hat{F}(x) = 1 - \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x - u}{\sigma}\right)^{-1/\hat{\xi}}$$

Here, given u, $\hat{\sigma}$ and $\hat{\xi}$ are estimated by maximum likelihood method. Due to the simple explicit form of the GPD, this estimation is used widely in calculating risk measures such as VaR and Expected Shortfall.

Case studies for insurance data using GPD fitting can be found in [McNeil, 1997] and [Resnick, 1997]. [Rootzén and Tajvidi, 1997] provided a full analysis using GPD on windstorm data. [Longin and Solnik, 2001] used a bivariate model in which each marginal tail is a GPD and the dependence structure is captured by a function from [Ledford and Tawn, 1997] to study stock market indices in different countries.

GPD has recently gained attention in literature on operational risk modeling. The most explored as well as most promising approach in this field of study is the Loss Distribution Approach (LDA) in which a frequency process and a severity process of losses are combined to estimate accumulated loss. For most severity distributions and frequency processes, a closed form for the aggregated loss does not exist. Instead, simulations and numerical techniques are used to arrive at any type of approximation. [Moscadelli, 2004] studied operational loss data collected by the Basel Committee in the year 2001. He concluded that GPD seemed to be good fit for the data with the tail index α between .7 and 1.7 for different business lines. Using this observation, [Bocker and Kluppelberg, 2005] investigated a simple LDA model in which the severity of the loss is GPD and is assumed to be independent from the frequency process of losses. They came up with a remarkably simple closed form approximation for OpVaR (Operational Value-at-Risk) which only required estimations of the GPD parameters and the expected value of the frequency process.

Another use of the GPD is in dynamic risk modeling. All the approximations

so far have assumed an i.i.d. dataset, which does not hold true for financial returns. [McNeil and Frey, 2000] proposed a method in which a GARCH stochastic volatility model is first fitted to historical data set. From this fitting, innovations are extracted. These innovations are more likely to be i.i.d, or at least, uncorrelated, than the original data. GPD is applied to estimate extreme quantile of the innovations and VaR of the returns can then be calculated.

1.3.3 Other parametric methods for modeling extreme data

In contrast to using a GPD, where only the extreme data is fitted, there are numerous other studies in which a heavy-tailed parametric distribution is fitted to the whole data set. The most commonly used distributions are the t-distribution and stable distribution.

For an extensive review on univariate and multivariate stable distribution, see [Samorodnitsky and Taqqu, 1994]. Numerical techniques related to stable distribution including simulation and parameter estimation can be found in [Adler et al., 1998]. The advantages of using a stable distribution, beside being able to capture both heavier tails and asymmetry seen in financial data, are discussed in [Mittnik et al., 1998]. A major reference for applications in finance of stable models, from option pricing to portfolio analysis is in [Rachev and Mittnik, 2000]. The use of stable distribution in VaR models have recently been studied by [Khindanova et al., 2001], [Mittnik et al., 2002], [Stoyanov et al., 2006]. [Rachev, 2003] provides an extensive collection of research focusing on the use of stable models in finance, ranging from calculating market and credit VaR to asset allocation.

The t-distribution is also a relatively simple distribution that can capture the heavier tails of financial data. However, in contrast to the stable family, there is no allowance for asymmetry in the standard t distribution. The most common use of t- distribution is to incorporate it into a GARCH type process. [Bollerslev, 1987] was the first to suggest this idea and applied this to foreign exchange rate returns. Other works of similar approach include [Baillie and Bollerslev, 1989], [Baillie and DeGennaro, 1990] and [Beine et al., 2002]. Extensions to capture skewness, called the skewed-t distributions, and their applications in finance have been studied by [Hansen, 1994], [Theodossiou, 1998], [Bauwens and Laurent, 2005] and [Aas and Haff, 2006] among several others.

1.4 Current methods of choosing extreme data

1.4.1 Extreme univariate data

Regardless of the method chosen to model risks, either semi parametric or parametric, there is a common problem of choosing a threshold. Most studies that have been done on the choice of extreme data have focused on improving the estimation of the tail index. This problem of choosing the tail is particularly critical because the Hill estimator has proved to be very sensitive to the choice of k. This sensitivity is shared by any other estimators of the tail index, such as the Pickands estimator and the moment estimator.

There is an inherent bias-variance tradeoff to tail index estimators when we choose k. The specific nature of this tradeoff depends on the specific choice of k and the underlying distribution F. To formalize this notion, de Haan and Peng [1998]

proved the following result using the concept of second order regular variation which we will state formally. Let $U = \left(\frac{1}{1-F}\right)^{\leftarrow}$ be the generalized inverse function. We assume that there exists $\rho < 0$ and a function A such that $|A|$ is regularly varying at infinity with exponent ρ, and such that for all $x > 0$

$$\lim_{r\to\infty} \frac{\frac{U(rx)}{U(r)} - x^{\gamma}}{A(r)} = x^{\gamma}\frac{x^{\rho}-1}{\rho} \tag{1.15}$$

(recall that $\gamma = 1/\alpha$). We also assume, without loss of generality, that A is continuous and $|A|$ is eventually decreasing.

Assuming the above second order condition holds, $n \to \infty$, $k \to \infty$ $\frac{k}{n} \to 0$ and $\sqrt{kA}\left(\frac{n}{k}\right) = C \in (-\infty, \infty)$ then de Haan and Peng [1998] proved that for the Hill estimator

$$\sqrt{k}(H_{k,n} - \gamma) \xrightarrow{d} N(\frac{C}{1-\rho}, \gamma^2)$$

This explains the phenomenon that the larger the choice of k, the smaller the variance and the bigger the bias.

To decide on a good k, sometimes visual techniques are used: the estimator, e.g. the Hill estimator, is plotted for a range of k, and then one looks for a part of the plot that looks stable. The corresponding stable value is used to estimate α, and several smoothing techniques have been introduced to assist in this visual analysis; see [Resnick and Stărică, 1997]. Still, the procedure is difficult to automate, and even visually it is sometimes difficult to use, as the so-called "Hill horror plots" demonstrate; see [Embrechts et al., 1997]. Examples of Hill plots are shown in Figure 1.2 and Figure 1.3. There also exists the equivalent plot for Pickands and the moment estimator. Attempts to improve the Hill plots included the smooHill plot, altHillplot (see [Resnick and Stărică, 1997; Drees et al., 2000]).

A systematic way of selecting a "good" number of upper order statistics in Hill

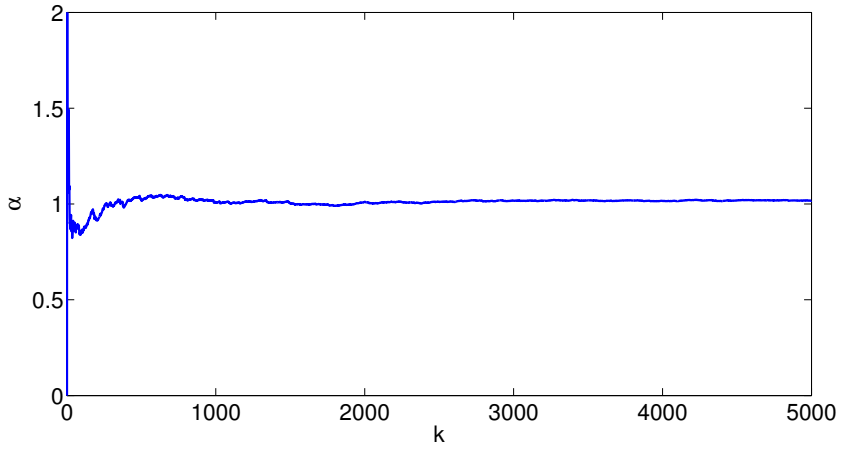

Figure 1.2: Hill plot for Pareto(1) with 5000 observations

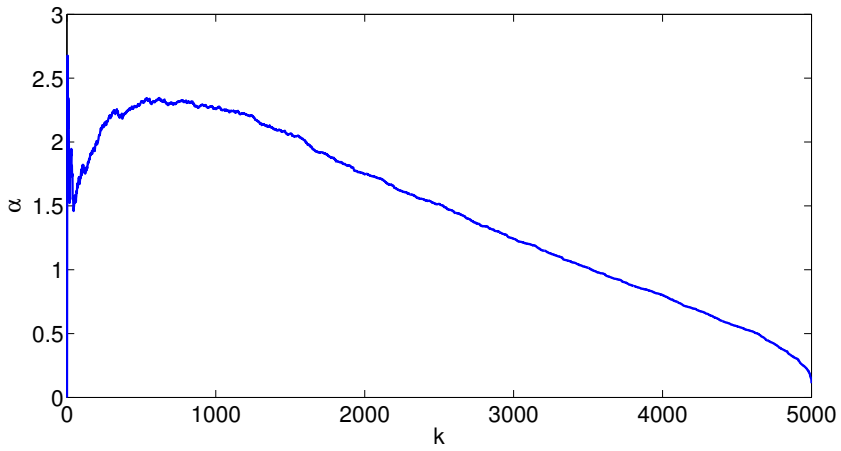

Figure 1.3: Hill "horror" plot for Stable(1.7) with 5000 observations

estimator originated with [Hall, 1990] and is based on the assumption that distribution F satisfies a further assumption of second order regular variation (which we introduce below). Under this assumption, it becomes possible to look for $k = k(n)$ that minimizes the asymptotic bias of the estimator. The method was later refined by [Danielsson et al., 2001], who suggested a two-level bootstrap procedure that works under minimal a priori available information. An alternative approach of finding this asymptotically optimal number k of upper order statistics was suggested, under slightly more restrictive assumptions, by [Drees and Kaufmann,

1998].

Hill himself suggested a sequential statistical procedure for choosing k in his original paper [Hill, 1975]. He considered the case when the distribution F had an exact Pareto tail beyond an unknown threshold D. If $X_{n-k-1,n} > D$, then, under the exact Pareto tail assumption, the random variables $iV_i := i \log \frac{X_{n-i,n}}{X_{n-i-1,n}}$ for $i = 0, 1, 2 \ldots, k$ are independent exponential random variables of parameter α. On the other hand, for k too large, the behavior of $\{iV_i\}$ would exhibit discrepancies from the exponential distribution. One can sequentially use exponential goodness of fit tests on $\{iV_i : i = 1, \ldots k\}$ for increasing k, until the hypothesis of exponentiality is rejected. [Hall and Welsh, 1985] argues that the procedure tends to result in too large a number k of order statistics.

1.4.2 Extreme multivariate data

The only existing systematic method to choose k appears to be the "Stărică plots," introduced in [Stărică, 1999]. It is a visual method, based on the fact that the tail measure ν_* has the scaling property (1.8). For bivariate data, define the set $A = \{(x, y) : x \geq 0, y \geq 0, ||(x, y)|| \geq 1\}$. Then by the scaling property

$$\nu_*(cA) = c^{-1}\nu_*(A),\ c > 0\,.$$

Thus, if we graph the function $c \mapsto \frac{\hat{\nu}_*(cA)}{c^{-1}\hat{\nu}_*(A)}$, for a good choice of k, this function would be approximately 1 for values of c near 1. This function is plotted for various values of k and checking is performed graphically until we find a k that would give us the right plot. One would then use the observations of the norm larger than the kth radial order statistic in estimating the spectral measure.

To be precise, if we use Equation (1.12) for the estimates of $\hat{\nu}_*(cA)$ and $\hat{\nu}_*(A)$,

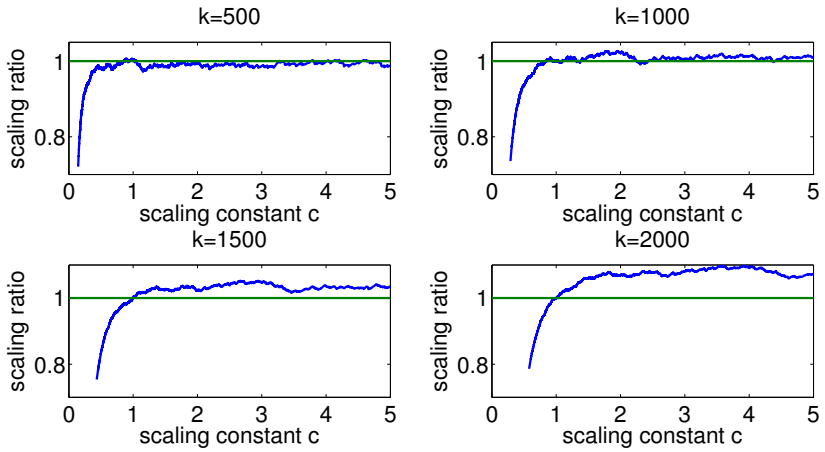

Figure 1.4: Stărică plots for bivariate Pareto with 5000 observations

the Stărică plot method is as follows. For a fixed k, we perform the rank transform in Equation 1.11 and let $c = R_{(j),k}$, the jth largest radius. Then the function we need to plot is

$$R_{(j),k} \mapsto \frac{R_{(j),k} \cdot j}{\sum_{i=1}^{n} 1_{\left(R_{i,k}>1\right)}},\ j = 1, ..., n$$

This method is demonstrated in the following example. Let X_1, X_2 be independent Pareto random variables with tail indice being 1 and 2 respectively. The sample size is 5000. The chosen norm is l^2. We plot the Stărică plots for different values of k in Figure 1.4.

From the plots, $k = 500$ is a little too low while $k = 1500$ or $k = 2000$ seems to be too much. This leaves us the choice of $k = 1000$. We can see that Stărică plots are useful in providing a rough guidance on k. An attempt to automate this process by using some distance measure between the plot and the horizontal line $y = 1$ proved to be unreliable ([Resnick, 2007]).

CHAPTER 2

A METHOD OF CHOOSING THE TAIL FOR UNIVARIATE DATA

2.1 The method

Instead of viewing the problem of selecting the number k in Hill estimator (or any related estimator) as a problem of optimizing asymptotic efficiency, we view it as the problem of deciding which part of a given sample contains reliable information on the tail of the distribution F. Put another way, we would like to know where in a sample the tail begins. Much of our motivation lies in the multivariate context: given a sample of random vectors (potentially, in a very high dimensional space) with an appropriately defined regularly varying tail we would like to test a variety of different subvectors of these vectors for tail independence. This involves repeated estimation of the so called tail measure, a time consuming procedure, which is also highly sensitive to the contamination of the tail by the center of the distribution (see [Resnick, 2007]). It is, therefore, desirable to have a reasonably quick way of deciding which part of the sample belongs to the tail.

Our approach is based on a simple idea which we now introduce informally. It is well known that, under the assumption (1.1) of regular variation, vague convergence of point processes holds,

$$N_n = \sum_{i=1}^{n} \delta_{X_i/a_n} \stackrel{v}{\Rightarrow} N_* \,,$$

where δ_x is a point mass at x, and (a_n) a positive sequence satisfying $\bar{F}(a_n) \sim 1/n$ as $n \to \infty$. Further, N_* is a Poisson random measure on $(0, \infty]$ with intensity measure $\mu_*(x, \infty] = x^{-\alpha},\ x > 0$; see [Resnick, 1987]. We interpret this result as saying that any upper order statistics in the sample that fall in the tail region

behave like points of a Poisson random measure with a power intensity. This property can be tested statistically, and sequentially. Specifically, one can perform appropriate statistical tests on the subsamples $X_{n-k+1,n}, X_{n-k+2,n}, \ldots, X_{n,n}$ while increasing k, and terminate the procedure once the k upper order statistics stop resembling points of a Poisson random measure with a power intensity.

We use the following property of a Poisson process: if $V_1 > V_2 > \ldots > V_k$ are the largest points of a Poisson process on $(0, \infty)$ with the mean measure $\mu_*(x, \infty) = x^{-\alpha}$, $x > 0$, then $\{V_i/V_k,\, i = 1, \ldots, k-1\}$, considered as a set, forms an i.i.d. sample from the Pareto distribution with the tail $x^{-\alpha}$, $x > 1$, and taking the logarithms transforms this sample into an i.i.d. sample of exponential random variable with the mean $\gamma = 1/\alpha$. Accordingly, our procedure for deciding on the number k of the upper order statistics to use in the Hill estimator consists of sequentially testing the samples $\{\log \frac{X_{n-i,n}}{X_{n-k,n}} : i = 0, 1 \ldots k-1\}$ for the null hypothesis of exponential distribution. Our choice of sample fraction k used in the tail estimation is then $N_n - 1$, where N_n is the smallest k such that the test described above rejects the null hypothesis of exponentiality.

In order to implement this procedure one has to choose a test of exponentiality. Once this has been done, the only remaining choice is that of the significance level of the test. Such choices are needed in all procedures to select the number of the order statistics to use (recall the subsample size in the bootstrap procedure of [Danielsson et al., 2001], or the threshold sequence of [Drees and Kaufmann, 1998]). We suggest a canonical way of choosing this significance level that appears to work reasonably well in the situations we have tried.

To test for exponentiality we choose the Greenwood-type statistic

$$Q_{k,n} = \frac{\sqrt{k}}{2}\left(\frac{\frac{1}{k}\sum_{i=0}^{k-1}\left(\log\frac{X_{n-i,n}}{X_{n-k,n}}\right)^2}{\left(\frac{1}{k}\sum_{i=0}^{k-1}\log\frac{X_{n-i,n}}{X_{n-k,n}}\right)^2} - 2\right);$$

Its large sample distribution under the null hypothesis assumption of exponentiality is the standard normal distribution ([Dahiya and Gurland, 1972]). Asymptotic behavior of this type of statistics in the context of extremes has been studied by [Neves and Fraga Alves, 2007] in a semi-parametric situation, and in a parametric situation by [Wang, 1995].

One could then try to implement a sequential testing procedure by choosing a critical value ω (perhaps, a 99% quantile with respect to the limiting standard normal distribution) and use $N_n^* - 1$ upper order statistics, where

$$N_n^* := \inf\{k : 1 \le k \le n,\, |Q_{k,n}| \ge \omega\} \tag{2.1}$$

The problem with this implementation is that N_n^* stays tight as the sample size increases, and this contradicts the obvious requirement that to get any averaging effect we need to take more and more order statistics into account. Therefore, the critical value needs to increase with the sample size n. We achieve this by selecting an increasing sequence $\theta_n \uparrow \infty$; this is the degree of freedom we mentioned above. On the other hand, in order to avoid taking into account too many order statistics, we choose to make it easier to reject the null hypothesis for larger k. It turns out that a good way to put all of this together is to set

$$N_n := \inf\left\{k : 1 \le k \le n,\, |Q_{k,n}| \ge \omega\sqrt{\frac{\theta_n}{k}}\right\}. \tag{2.2}$$

We will see in Theorem 1 below that, under a suitable growth condition on θ_n, this definition of N_n makes it, roughly, proportional to θ_n.

2.2 Results

To introduce our results, we required the assumption of second order regular variation. Let us restate this condition. Let $U = \left(\frac{1}{1-F}\right)^{\leftarrow}$ be the generalized inverse function. We assume that there exists $\rho < 0$ and a function A such that $|A|$ is regularly varying at infinity with exponent ρ, and such that for all $x > 0$

$$\lim_{r\to\infty} \frac{\frac{U(rx)}{U(r)} - x^{\gamma}}{A(r)} = x^{\gamma}\frac{x^{\rho}-1}{\rho} \tag{2.3}$$

(recall that $\gamma = 1/\alpha$). We also assume, without loss of generality, that A is continuous and $|A|$ is eventually decreasing. Under regular variation and second regular variation conditions, the following holds.

Theorem 1. *Let $\omega > 0$ and (θ_n) an increasing sequence such that $\theta_n \uparrow \infty$ and $\theta_n = \mathrm{o}\left(n^{\frac{2|\rho|}{1+2|\rho|}}\right)$ as $n \to \infty$. Then $\dfrac{N_n}{\theta_n} \Rightarrow \tau_\omega$, where τ_ω is the first time a standard Brownian motion hits $\pm\omega$.*

To prove the theorem, first we need the following lemma, which is a functional version of Lemma 3.5.5 in [de Haan and Ferreira, 2006]. In this lemma we work with spaces of the type $D[0,\infty)$, $D^2[0,\infty)$, $D[\delta,\infty)$ and $D^2[\delta,\infty)$ for $\delta > 0$. We endow the D^2 spaces with the (strong) J_1 topology. See [Whitt, 2002] for details.

Lemma 2.2.1. *Assume (θ_n) is an increasing sequence such that $\theta_n \uparrow \infty$ and $\theta_n = \mathrm{o}\left(n^{\frac{2|\rho|}{1+2|\rho|}}\right)$ as $n \to \infty$. For $n \geq 1$ define*

$$M^j_{\theta_n,n}(t) = \begin{cases} 0 & \text{if } \ 0 \leq t < \dfrac{1}{\theta_n} \\ \dfrac{1}{\lfloor \theta_n t \rfloor} \sum_{i=0}^{\lfloor \theta_n t \rfloor - 1} \left(\log \dfrac{X_{n-i,n}}{X_{n-\lfloor \theta_n t \rfloor, n}} \right)^j & \text{if } \ \dfrac{1}{\theta_n} \leq t \leq \dfrac{n}{\theta_n} \\ \dfrac{1}{n} \sum_{i=0}^{n-1} \left(\log \dfrac{X_{n-i,n}}{X_{1,n}} \right)^j & \text{if } \ t > \dfrac{n}{\theta_n}\,, \end{cases}$$

$j = 1, 2$. *Then*

$$\sqrt{\theta_n}t \begin{pmatrix} \dfrac{M^1_{\theta_n,n}(t)}{\gamma} - 1 \\ \dfrac{M^2_{\theta_n,n}(t)}{\gamma^2} - 2 \end{pmatrix} \Rightarrow \begin{bmatrix} W_1(t) \\ W_2(t) \end{bmatrix}$$

in $D^2[0,\infty)$*, where* $\big((W_1(t), W_2(t)),\, t \geq 0\big)$ *is a two-dimensional zero mean Brownian motion with covariance matrix*

$$\begin{bmatrix} 1 & 4 \\ 4 & 20 \end{bmatrix}.$$

Proof of Lemma. See Appendix A.1 □

Proof of Theorem 1. See Appendix A.2 □

The following theorem, which is the main theorem of this section, shows that using the Hill estimator with the random number of upper order statistics given by (3.3) is, indeed, a consistent estimator of $\gamma = 1/\alpha$. We also derive a weak limit result for the deviations of the estimator from $1/\gamma$. It shows that these deviations are of the order $1/\sqrt{\theta_n}$, which is expected, since by Theorem 1, the number N_n of the order statistics in the Hill estimator is of the order θ_n.

Theorem 2. *Let* $\theta_n = \mathrm{o}\left(n^{\frac{2|\rho|}{1+2|\rho|}}\right)$ *as* $n \to \infty$*, and let* N_n *be given by* (3.3)*. The Hill estimator based on* N_n *upper order statistics is consistent, i.e.*

$$H_{N_n,n} = \frac{1}{N_n} \sum_{i=0}^{N_n-1} \log \frac{X_{n-i,n}}{X_{n-N_n,n}} \xrightarrow{P} \gamma$$

as $n \to \infty$*. Furthermore,*

$$\sqrt{\theta_n} \left(\frac{H_{N_n,n}}{\gamma} - 1 \right) \Rightarrow \frac{G}{(\tau_\omega)^{1/2}}\,, \tag{2.4}$$

where G *is a standard normal random variable independent of the first hitting time* τ_ω *from Theorem 1.*

Proof. See Appendix A.3 □

Remark 3. We have already mentioned that, according to Theorem 2, the deviations of our estimator from the true value of γ are of the order $1/\sqrt{\theta_n}$. Since, under conditions of that theorem, the rate of growth of θ_n can go all the way up to $n^{\frac{2|\rho|}{1+2|\rho|}}$, our estimator can almost achieve the optimal rate of decay of the asymptotic deviation from the true γ, given by $n^{\frac{-|\rho|}{1+2|\rho|}}$; see e.g. [Danielsson et al., 2001]. Since the exponent ρ of the second order regular variation (2.3) is unknown, one could, potentially, combine our method with the bootstrap technique of [Danielsson et al., 2001]. We do not pursue this approach. In fact, our goal is not necessarily asymptotic efficiency, since for some distributions it can take a very long time until these asymptotics become effective. Our goal is, rather, determining, for a given sample size, which (upper) part of the sample appears, statistically, to be consistent with being in the tail region. For this purpose, sequences (θ_n) that increase at a much slower rate, appear to be appropriate. In fact, as the reader will see in the subsequent sections, we advocate using logarithmically fast increasing sequences.

2.3 Simulation results

In this section we evaluate our procedure (3.3) for selecting the number of upper order statistics in the Hill estimator on simulated univariate data. We compare the resulting performance of the estimator with the bootstrap procedure of [Danielsson et al., 2001] , the optimal sample fraction choice of [Drees and Kaufmann, 1998], and to the original testing procedure of [Hill, 1975]. For the test data we choose i.i.d. samples from the Student-t distribution, the Burr distribution and the symmetric stable distribution. We have chosen these distributions because the "usual" Hill plots are often difficult to interpret in these cases. Recall that for the

Student-t distribution, the tail exponent α is equal to the number of degrees of freedom. For the Student-t distribution we considered the cases $\alpha = 1, 3, 4$, while for the stable distribution, we tested the cases $\alpha = 1$ and $\alpha = 1.7$. For the Burr distribution, we chose Burr(1,-.5) and Burr(1,-2) where the distribution function of Burr(α, ρ) is as follow

$$\bar{F}(x) = (1 + x^{-\rho * \alpha})^{\frac{1}{\rho}}$$

The Hill estimator has been shown to be consistent not only on i.i.d. data but also under certain kinds of dependence, (see [Hsing, 1991], [Resnick and Stărică, 1995], [Resnick and Stărică, 1998]). This includes the class of moving average processes, and we also test our estimator on the MA(1) process Y_t=$X_t + X_{t-1}$ where X_t are i.i.d. Student-t random variables with 3 degrees of freedom. The tail index in this case is equal to 3.

We have tested our estimator with logarithmic sequences $\theta_n = \log n$ and $\theta_n = (\log n)^2$. In all cases we chose ω to be the 95th quantile of the standard normal distribution. The results of the simulation are displayed in the tables below. Table 1 and 2 present the results using sample size $n = 5000$ and $n = 50000$ of absolute values of the above distribution respectively. Each simulation was performed 250 times. The following information is displayed.

- The testing procedure suggested by [Hill, 1975] using the moment statistic with significant level .05.
- Our choice of the sample fraction N_n with $\theta_n = \log n$.
- Our choice of the sample fraction N_n with $\theta_n = (\log n)^2$.
- The choice of sample fraction using bootstrap method proposed by [Danielsson et al., 2001] with n_1 going from 1000 to 4000 in increments of 250 for

sample size 5000 and from 10000 to 35000 in increments of 2500 for sample size 50000. In both case the number of bootstrap samples is 500.

- The optimal sample fraction choice $\hat{k}_{opt}$ of [Drees and Kaufmann, 1998] with the initial $\tilde{\beta}_n$ based on the upper $2\sqrt{n}$ order statistics, $r_n = 2.5\tilde{\beta}_n n^{.25}$, $\psi = .7$ and $\rho_0 = 1$ (in their terminology ρ is a positive number).

The results in Table 1 and Table 2 indicate that we consistently obtain good results with $\theta_n = (\log n)^2$. Our choice of the number of order statistics to use in the Hill estimator performs reasonably well with both moderate and large sample sizes. It compares particularly favorably with other methods when the sample size is moderate and tail estimation is difficult (as the notoriously difficult case of the stable distribution with $\alpha = 1.7$ shows.) On the other hand, when many methods do well, so does our approach and, in that case, the root mean square error obtained with our method may be larger than that produced by other methods. However, *all* mean square errors are then small, and our method, along with methods, still estimates α well.

Our method is very computationally efficient, and is easy to automate. This is particularly important in financial applications where regular estimation of the tail index of around $15,000$ securities based on around $2,000$ observations is not uncommon.

We finish this section with a number of additional comments.

- It follows from Theorem 1 and our choice of logarithmically increasing sequence (θ_n), that our method tends to result in a smaller k than, say, the $\hat{k}_{opt}$ method. For example, in our simulation of samples of 50000 Student-t random variables with degree of freedom 4, the average number of order

statistics N_n used by our method with $\theta_n = (\log n)^2$ is 320, whereas the $\hat{k}_{opt}$ method choose on average 500 data points.

- None of the methods does particularly well when applied to a sample that does not have a regularly varying tail. We have tried these approaches on samples of 5000 standard lognormal random variables. On average, between 30 and 800 data points are taken to be in the tail, with our method ($\theta_n = (\log n)^2$) averaging at 130. This results in estimated values of α between 2 and 4. With the choice of $\theta_n = (\log n)^2$, the mean estimate is 2.8.

- If the underlying distribution is very close to Pareto (for example, it is exactly Pareto after a threshold), the sequential testing essence of our method ensures that one of the tests will fail, and the procedure stops, well before the threshold is reached (the tail is still estimated well). of course, the exponent ρ of the second order regular variation will be very large in absolute value in this case. If this is known ahead of time, then the sequence θ_n can be taken to grow almost linearly fast, which will result in a large part of the sample taken into account. In general, if ρ is known ahead of time, we can allow θ_n to grow faster, and this usually results in a smaller root mean square error. We simulate a distribution where $\frac{2}{3}$ of the data is truncated standard normal on the interval $[0, 1]$ and $\frac{1}{3}$ of the data is pure Pareto(1). Hence, the tail starts at 1 and comprises exactly of $\frac{1}{3}$ of the data. We let $\theta_n = n^{.9}$. For sample size of 5000, out of 250 simulations, N_n averages at 1770 while with sample size of 50000, mean N_n is 16200. The mean estimates of α are 1.0426 and 1.0111 respectively.

Table 2.1: Simulation with $n = 5000$

Method			Hill		N_n, $\theta_n = \log n$		N_n, $\theta_n = (\log n)^2$		Bootstrap		$\hat{k}_{opt}$	
Dist.	α	$\lvert\rho\rvert$	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
Student(4)	4	.5	2.7794	1.7098	4.0496	.9128	3.4568	.6510	3.6135	.6859	3.4270	.6629
Student(3)	3	.667	2.1719	.9843	3.1460	.7434	2.7726	.3657	2.8490	.4383	2.7669	.3358
Student(1)	1	2	.9326	.3937	1.0738	.2203	1.0109	.0890	.9881	.0502	.9965	.0391
Stable(1.7)	1.7	2	2.0347	.6951	1.9401	.4586	2.0013	.3887	2.2515	.5654	2.2138	.5283
Stable(1)	1	2	.8965	.1683	1.0684	.2567	1.0099	.0855	.9945	.0689	.9912	.0404
Burr(1,-.5)	1	1	.7457	.5154	1.0459	.2702	.9009	.1425	.9036	.1666	.8772	.1431
Burr(1,-2)	1	1	.9140	.1658	1.0922	.2528	1.0178	.0844	.9906	.0477	.9910	.0742
MA(1)	3	.667	2.8335	1.8239	3.8765	1.6059	3.1434	.5232	3.1365	.5647	3.0955	.3708

Table 2.2: Simulation with $n = 50000$

Method			Hill		N_n, $\theta_n = \log n$		N_n, $\theta_n = (\log n)^2$		Bootstrap		$\hat{k}_{opt}$	
Dist.	α	$\lvert\rho\rvert$	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
Student(4)	4	.5	3.2954	1.5064	4.2361	1.0055	3.7958	.4743	3.7690	.5282	3.6080	.4217
Student(3)	3	.667	2.5280	.6734	3.1677	.7360	2.9391	.2245	2.8900	.3013	2.8490	.1839
Student(1)	1	2	.9499	.2182	1.0857	.2382	1.0103	.0697	.9959	.0215	.9970	.0159
Stable(1.7)	1.7	2	2.0894	.5303	1.8276	.3935	1.7733	.1670	2.2276	.5288	2.1057	.4079
Stable(1)	1	2	.9608	.1357	1.0480	.2022	1.0079	.0764	.9918	.0204	.9965	.0165
Burr(1,-.5)	1	1	.8222	.2730	1.0437	.2166	.9593	.0762	.9435	.1113	.9280	.0833
Burr(1,-2)	1	1	.9588	.1082	1.0588	.2736	1.0144	.0724	.9930	.0173	.9950	.0149
MA(1)	3	.667	3.6480	5.4884	3.5814	1.2982	3.1893	.4743	3.1014	.2113	3.0898	.1775

CHAPTER 3

A METHOD OF CHOOSING THE TAIL FOR MULTIVARIATE DATA

3.1 The method

Recall that to estimate the tail measure by the rank transform method, given a d-dimensional sample of size n, for each $j = 1, \ldots, d$, we use the following result

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{\left(\frac{k}{r_i^{(j)}}, j=1,\ldots,d\right)} \Rightarrow \nu_* \tag{3.1}$$

where

$$r_i^j = \sum_{m=1}^{n} 1_{[Z_m^{(j)} \geq Z_i^{(j)}]}, \quad 1 \leq i \leq n\,,$$

which is the sequence of ranks and $k = k(n)$ is an intermediate sequence. We then apply the polar transformation on the rank data. For the rest of this chapter, we use the l^∞ unit sphere for our examples and the polar transformation for nonnegative bivariate data is defined as

$$(R_{i,k}, \theta_{i,k}) = \left(\left\| \left(\frac{k}{r_i^{(j)}}, j = 1, 2 \right) \right\|_\infty, \; arctan \frac{\frac{k}{r_i^{(2)}}}{\frac{k}{r_i^{(1)}}} \right), \; i = 1, \ldots, n\,.$$

This then gives us the following consistent estimator for the spectral measure S_*:

$$S_{*k,n}(\cdot) := \frac{\sum_{i=1}^{n} 1_{\left(R_{i,k}>1,\, \theta_{i,k} \in (\cdot)\right)}}{\sum_{i=1}^{n} 1_{\left(R_{i,k}>1\right)}} \Rightarrow S_*(\cdot)\,. \tag{3.2}$$

The remaining problem is to decide on the choice of k to use in (3.1) and (3.2). We now proceed to extend our sequential testing procedure to pick an appropriate k in univariate to the multivariate framework of estimating the spectral measure.

Specifically, we start by applying the testing procedure to each set of marginal observations. For the jth marginal, $j = 1, \ldots, d$ we calculate the first rejection time, $N^{(j)}$, via

$$N_n^{(j)} := \inf \left\{ k : 1 \leq k \leq n,\, |Q_{k,n}| \geq \omega_j \sqrt{\frac{\lambda_n^{(j)}}{k}} \right\} , \tag{3.3}$$

using the original data for that marginal, with some $\omega_j > 0$ and some sequence $(\lambda_n^{(j)})$. Any choice of (ω_j) is possible; again, we can fix them and find appropriate $(\lambda_n^{(j)})$. We set our choice of k to be

$$N_n = \bigwedge_{j=1}^{d} N_n^{(j)} . \tag{3.4}$$

Taking the minimum in (3.4) means that our procedure is, generally, conservative about deciding on the "tail part" of the data. Other choices of k are possible, with the largest rejection point, $N_n = \bigvee_{j=1}^{d} N_n^{(j)}$, being an obvious choice. When tested on simulated data, the latter choice of k often appeared to be "too generous" with deciding which observations are "in the tails".

3.2 Results

The consistency of the resulting estimator of the tail measure is proven in the following theorem.

Theorem 4. *Assume that the marginal and the joint regular variation conditions* (1.3) *and* (1.4) *hold. Assume, further, that the jth marginal distribution satisfies the second order regular variation condition with exponent $\rho_j < 0$, $j = 1, \ldots, d$. Let $\omega_j > 0$ and $(\lambda_n^{(j)})$ be an increasing to infinity sequence such that*

$\lambda_n^{(j)} = \mathrm{o}\left(n^{\frac{2|\rho_j|}{1+2|\rho_j|}}\right)$, $j = 1, \ldots, d$, *and let* N_n *be defined by* (3.4). *Then*

$$\frac{1}{N_n}\sum_{i=1}^{n} \epsilon_{\left(\frac{N_n}{r_i^{(j)}}, j=1,\ldots,d\right)} \Rightarrow \nu_*$$

weakly in the vague topology.

Proof. See Appendix A.4 □

Remark 5. It is important to note that our consistency result does require the second-order regular variation assumption, unlike most of the previous approaches, in which only the (first-order) regular variation is used. When a non-random number k of upper order statistics is used in estimation, the second-order regular variation is often used to obtain asymptotic normality of the estimator; see e.g. [de Haan and Ferreira, 2006]. With a random choice of k as in our procedure, we needed to use the second-order regular variation even to obtain consistency of the estimator, even in the one-dimensional case. In fact, the behaviour of the random variable N_n above seems to undergo a phase transition if the condition $\lambda_n = \mathrm{o}\left(n^{\frac{2|\rho|}{1+2|\rho|}}\right)$ breaks down. It remains to be understood what the behaviour of our estimator is if this happens (or if the assumption of the second-order regular variation does not hold at all). However, as we will see in the sequel, the choice of the sequences $(\lambda_n^{(j)})$ we advocate (as leading to a good performance) is logarithmic and, hence, does not require knowing the value of the exponent of the second-order regular variation.

The following is an immediate corollary of Theorem 1. It provides the estimator of the spectral measure we will use in the rest of the paper.

Corollary 6. *Under the assumptions of Theorem 1,*

$$S_{*N,n}(\cdot) = \frac{\sum_{i=1}^{n} 1_{\left(R_{i,N}>1,\, \theta_{i,N}\in(\cdot)\right)}}{\sum_{i=1}^{n} 1_{\left(R_{i,N}>1\right)}} \Rightarrow S_*(\cdot)\,, \tag{3.5}$$

weakly on the unit sphere.

Note that this result is a weak convergence result. Other procedures have been shown to lead, under appropriate assumptions, to uniform convergence over certain classes of sets; see e.g. [Einmahl et al., 1993] and [Einmahl et al., 2001].

3.3 Simulation results

In this section, we evaluate our estimator (3.5) of the spectral measure on simulated data. We consider 3 two-dimensional examples. The first example looks at two scenarios with asymptotic independence (so that the true spectral measure is concentrated at the extreme points of the arc), the second example looks at a case of a complete tail dependence (so that the true spectral measure is concentrated at a single point in the interior of the arc), and the last example considers a situation where the true spectral measure is absolutely continuous with respect to the Lebesgue measure on the arc. In each case we ran simulations with sample sizes $n = 1000$, $n = 5000$ and $n = 20000$, and each simulation was performed 500 times. The results we report are the averages obtained from the 500 simulations. In all cases, we choose the numbers $N^{(j)}$, $j = 1, 2$ by (3.3) with $\lambda_n^{(j)} = (\log n)^2$, as recommended in the previous chapter. This choice of λ_n was empirically observed to work well in the univariate estimation of the exponent of regular variation for a variety of models. It has an added advantage of satisfying the condition $\lambda_n = \mathrm{o}\left(n^{\frac{2|\rho|}{1+2|\rho|}}\right)$ regardless of the actual value of the exponent ρ in the second-order regular variation. We also choose $\omega_j = 1.64$ for all j. This is the rejection level of the original exponentiality test (Dahiya and Gurland [1972]) with significance level .1, as in the previous chapter.

Example 3.3.1. The generic model is $X^{(i)} = |Z + Y^{(i)}|, i = 1, 2$ where $Z \sim N(0, 1)$ and $Y^{(1)}$ and $Y^{(2)}$ are independent identically distributed random variables independent of Z, with regularly varying tails.

1. $Y^{(i)}$, $i = 1, 2$ have a t distribution with 3 degrees of freedom. In this case, the second order regular variation index is $\rho = -\frac{2}{3}$. A simple way to calculate ρ based on the power expansion of the generalized inverse function u is given in Beirlant et al. [2004].

2. $Y^{(i)}$, $i = 1, 2$ have a Generalized Pareto distribution with parameters $\mu = 1, \alpha = 1, \sigma = 2$. Recall that the distribution function of such a distribution is given by
$$G_{\alpha,\sigma}(y) = 1 - \left(1 + \frac{1}{\alpha}\frac{(y-\mu)}{\sigma}\right)^{-\alpha}, \quad y \geq \mu\,. \tag{3.6}$$
In this case, $\rho = -1$ (regardless of what the location parameter μ is).

Our decision on the tail part of the sample plays a role not only in the estimation of the spectral measure, but also in estimation of the parameters of the distribution. Estimating the latter is not necessary for the estimation of the spectral measure, which is our main goal. However, the interplay between estimation of the parameters and the choice of the part of the data to use for it, is instructive, and we include a short discussion of the resulting insights.

In Example 3.3.1 (a) we estimate the marginal tail index (coinciding with the number of the degrees of freedom) by the Hill estimator, which is defined as follows. If $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ are the order statistics from a positive sample, $X_1, \ldots, X_n$, then the Hill estimator based on k upper order statistics is
$$H_{k,n} := \frac{1}{k}\sum_{i=0}^{k-1} \log \frac{X_{n-i,n}}{X_{n-k,n}}\,. \tag{3.7}$$

If the observations are i.i.d., with a regularly varying right tail with tail index α, then, with $n \to \infty$, $k \to \infty$ and $\frac{k}{n} \to 0$, the Hill estimator $H_{k,n}$ converges in probability to $\gamma = \frac{1}{\alpha}$; see Mason [1982]), who also shows, under certain conditions of $k = k_n$, equivalence between consistency of the Hill estimator and regular variation. We follow the previous chapter in our choice of k in estimating the jth marginal tail index. We choose it to be $N_n^{(j)}$ as above, $j = 1, 2$.

In the case of the generalized Pareto distribution of Example 3.3.1 (b) we use the $N_n^{(j)}$th largest order statistic of X_j as the location parameter $\mu^{(j)}$, $j = 1, 2$. The reason for this is that a generalized Pareto distribution is used as a model for the "excess tails" once a certain threshold has been crossed, for data in which the "real distribution" is not known; this is the so called POT (peaks over threshold) method; see e.g. Embrechts et al. [1997]. Choosing the threshold is often difficult, as it is supposed to represent a point beyond which the "tail behaviour" of the distribution is observed. This fits well into our approach, where $N_n^{(j)}$th largest order statistic represents the estimated "beginning of the tails". Therefore, instead of using the maximum likelihood estimation to estimate all the parameters of the known distribution in this example, we prefer to set the initial point $\mu^{(j)}$ of the generalized Pareto distribution, by what the data indicates, i.e. by the $N_n^{(j)}$th largest order statistic. Once the location parameter is decided on, the remaining parameters α and σ are obtained via maximum likelihood estimation.

We note that in both parts of Example 3.3.1 the tails are asymptotically independent, and the spectral measure splits its mass equally between the two extreme points, at the angles 0 and $\pi/2$.

Our general format for reporting the results of estimation is as follows. We report the numbers $N_n^{(j)}$, $j = 1, 2$, averaged over all the runs, as well as their

n	$\alpha^{(1)}$	$\alpha^{(2)}$	$N^{(1)}$	$N^{(2)}$	N
1000	2.9906	2.9819	111.8	110.3	82.3
5000	3.1411	3.1174	204.8	210.3	131.4
20000	3.1011	3.1217	330.8	299.7	195.9

Table 3.1: Simulation results for Example 3.3.1 (a)

n	Adaptive bias	Adaptive variance	5% bias	5% variance
1000	.0337	.0026	.0246	$4.784 * 10^{-4}$
5000	.0127	.0016	.0243	$9.836 * 10^{-5}$
20000	.0065	$6.519 * 10^{-4}$	.0243	$2.433 * 10^{-5}$

Table 3.2: Simulation results for Example 3.3.1 (a)

minimum, N_n, also averaged over all the runs. To measure how close the estimated spectral measure is to the true one, we calculate and report the integrated squared bias and variance of the estimator, given, respectively, by $\int_0^{\pi/2}\big(S_*([0,\theta]) - E(\hat{S}_*([0,\theta]))\big)^2\,d\theta$ and $\int_0^{\pi/2} Var\big(\hat{S}_*([0,\theta])\big)\,d\theta$. Here $S_*([0,\theta])$ is the mass assigned by the true spectral measure S to the part of the l^∞ unit sphere whose angle, in the polar coordinates, is in the interval $[0,\theta]$, $0 \le \theta \le \pi/2$. Similarly, $\hat{S}_*([0,\theta])$ is the mass assigned by the estimated spectral measure to the same set; these are the "cumulative distribution functions" indexed by the angle. The expected value and variance in the integrals are replaced by sample mean and sample variance. We compare the calculated biases and variances obtained using our approach to the "rule of thumb" choice of the part of the sample whose radial component is in the top 5% among all the observations. The alternative "Stărică plot" is not useful in this situation; it cannot distinguish between different k in the relevant range.

The estimation results for the t distribution of Example 3.3.1 (a) are reported in Table 3.1 and Table 3.2. We present estimates of the spectral measure from 20 simulations in Figure 3.1.

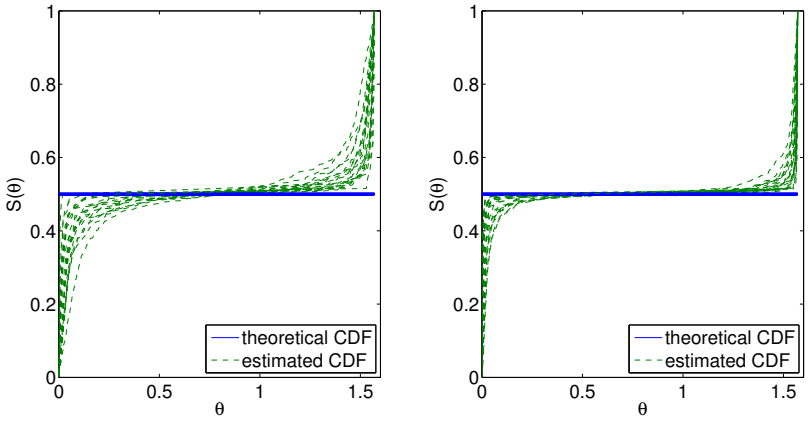

Figure 3.1: Sample spectral measures from 20 simulations of Example 3.3.1 (a) with $n = 5000$ and $n = 20000$

n	$N^{(1)}$	$N^{(2)}$	N
1000	153.9	148	96.9
5000	242.9	226	138.1
20000	315.7	313.1	192.7

Table 3.3: Simulation results for Example 3.3.1 (b)

The estimation results for the Generalized Pareto distribution of Example 3.3.1 (b) are similarly reported in Table 3.3, Table 3.4, Table 3.5 and Figure 3.2.

Observe that, while the "5% rule of thumb" results in a lower bias and variance than our estimator for the smallest sample size $n = 1000$, the situation is reversed for larger sample sizes. By construction, the bias of the 5% rule stays constant as sample size increases. On the other hand, our adaptive method quickly picks up the emerging tails and improves the quality of the estimation as the sample size increases. Our adaptive method needs some data to learn from and improve.

n	Adaptive bias	Adaptive variance	5% bias	5% variance
1000	.0328	.0037	.0217	$3.886 * 10^{-4}$
5000	.0119	.0015	.0214	$7.889 * 10^{-5}$
20000	.0062	$5.282 * 10^{-4}$	.0214	$2.026 * 10^{-5}$

Table 3.4: Simulation results for Example 3.3.1 (b)

n	$\alpha^{(1)}$	$\alpha^{(2)}$	$\sigma^{(1)}$	$\sigma^{(2)}$	$\mu^{(1)}$	$\mu^{(2)}$
1000	.9948	.9992	20.4228	22.2245	19.6233	22.5595
5000	1.0209	1.0042	70.3615	76.6433	77.2565	79.6734
20000	1.0172	1.0198	218.7182	210.5060	237.6335	231.3167

Table 3.5: Simulation results for Example 3.3.1 (b)

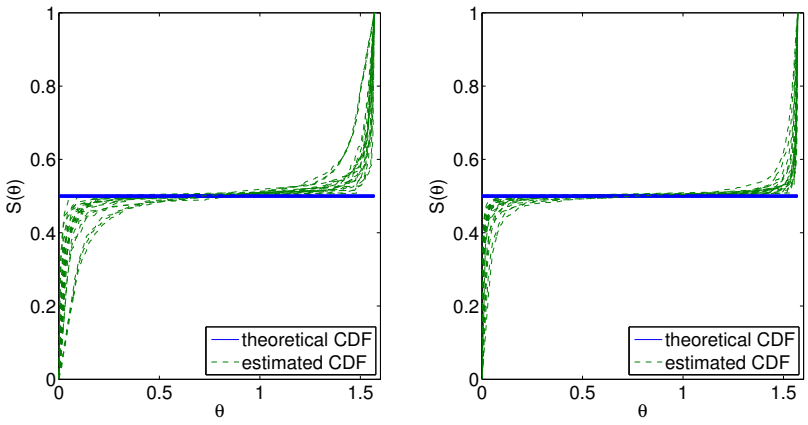

Figure 3.2: Estimated spectral measures from 20 simulations of Example 3.3.1 (b) with $n = 5000$ and $n = 20000$

As far as the estimation of other parameters is concerned, we see that we are estimating the tail exponent α well even for small sample sizes. However, we are missing the true values of the location parameter μ and the scale parameter σ of the Generalized Pareto distribution of Example 3.3.1 (b) by a wide margin. The reason, as we explained above, is that in deciding the location parameter we pretend that we do not not know the parametric form of the model and set the location parameter by the appropriate order statistic, similar to the POT method. This results in inflation of the location parameter. Recall that, in a Generalized Pareto distributed with parameters (μ, α, σ), given that $X > \mu + \delta$ for some $\delta > 0$, X has again a Generalized Pareto distribution, this time with parameters $(\mu + \delta, \alpha, \sigma + \delta/\alpha)$. Hence inflating the location parameter leads to overestimating the scale parameter as well. Note, however, that the tail exponent α is still adequately estimated.

n	$\alpha^{(1)}$	$\alpha^{(2)}$	$N^{(1)}$	$N^{(2)}$	N
1000	2.5393	1.0847	127.9	151.2	100.2
5000	2.2688	1.0277	199.1	243	165.4
20000	2.1528	1.0168	285.2	309	230.1

Table 3.6: Simulation results for Example 3.3.2

n	Adaptive bias	Adaptive variance	5% bias	5% variance
1000	.0384	.0031	.0248	.0014
5000	.0184	.0012	.0261	$2.821 * 10^{-4}$
20000	.0099	$5.907 * 10^{-4}$	.0261	$7.525 * 10^{-5}$

Table 3.7: Simulation results for Example 3.3.2

Example 3.3.2. This is an example of a bivariate law with a complete tail dependence. Let Y be Pareto(2) distributed, independent of two i.i.d. standard normal random variables $Z^{(1)}$ and $Z^{(2)}$. Let $X^{(1)} = Z^{(1)} + Y$ and $X^{(2)} = Z^{(2)} + Y^2$. The true spectral measure in this case has mass only at a single point, $\pi/4$. We present the results in Table 3.6, Table 3.7 and Figure 3.3. The lessons we learn here are similar to the lessons learned in Example 3.3.1; for larger sample sizes our algorithm estimates the spectral measure more precisely than the 5% rule does. Again, the reason here is that the bias of the 5% method remains constant as sample size increases, while our algorithm detects emerging tails for larger sample sizes.

Example 3.3.3. In this example the true spectral measure is absolutely continuous with respect to the Lebesgue measure on the l^∞ unit sphere. We consider two cases.

1. $X^{(i)} = Y^{(i)}$, $i = 1, 2$ with $\left(Y^{(1)}, Y^{(2)}\right)$ having the bivariate Cauchy distribution restricted to $(0, \infty)^2$. This distribution has the density given by

$$f(y^{(1)}, y^{(2)}) = \frac{2}{\pi}(1 + (y^{(1)})^2 + (y^{(2)})^2)^{-\frac{3}{2}}. \tag{3.8}$$

2. $X^{(i)} = Y^{(i)} + Z^{(i)}$, $i = 1, 2$, with $\left(Y^{(1)}, Y^{(2)}\right)$ having the bivariate Cauchy

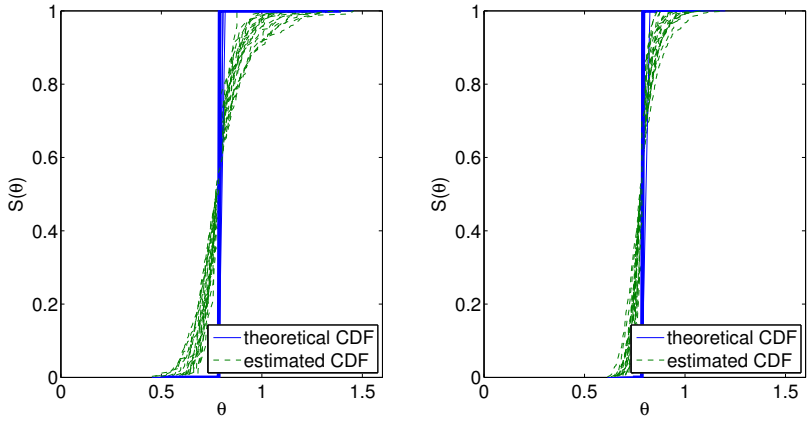

Figure 3.3: Estimated spectral measures from 20 simulations of Example 3.3.2 with $n = 5000$ and $n = 20000$

distribution restricted to $(0, \infty)^2$, independent of i.i.d. Pareto(2) random variables $Z^{(1)}$ and $Z^{(2)}$.

Note that, in both cases, the "c.d.f." of the true spectral measure on the unit sphere with respect to the l_p norm is : $S_*([0, \theta]) = \frac{1}{\sqrt{2}} \int_0^\theta ||\sin(\psi), \cos(\psi)||_p d\psi$, $0 \leq \theta \leq \pi/2$; it is, therefore, absolutely continuous with respect to the Lebesgue measure for any $1 \leq p \leq \infty$. It is, actually uniform (i.e. coincides, up to a multiplicative constant, with the Lebesgue measure) if $p = 2$. Recall that we are using the l^∞ norm for our estimation. The marginal distribution is Cauchy hence the exponent of the second-order regular variation is $\rho = -2$.

The results for part (a) of the example are reported in Table 3.8, Table 3.9 and Figure 3.4, and for part (b) of the example in Table 3.10, Table 3.11 and Figure 3.5.

We see that, as expected, it is harder to estimate the spectral measure in the "contaminated" case (b). Nevertheless, our approach works reasonably well. It is instructive to observe that in the pure bivariate Cauchy case of part (a) the

n	$\alpha^{(1)}$	$\alpha^{(2)}$	$N^{(1)}$	$N^{(2)}$	$\bar{N}$
1000	0.9814	1.0044	141.5	142.1	98.3
5000	1.0010	.9998	229.9	238.1	156
20000	1.0110	1.0067	301.8	302.5	201.2

Table 3.8: Simulation results for Example 3.3.3 (a)

n	Adaptive bias	Adaptive variance	5% bias	5% variance
1000	$1.876 * 10^{-4}$	.0016	$1.003 * 10^{-4}$	.0019
5000	$3.802 * 10^{-5}$	$9.692 * 10^{-4}$	$2.035 * 10^{-5}$	$3.883 * 10^{-4}$
20000	$1.767 * 10^{-5}$	$8.247 * 10^{-4}$	$2.023 * 10^{-5}$	$1.065 * 10^{-4}$

Table 3.9: Simulation results for Example 3.3.3(a)

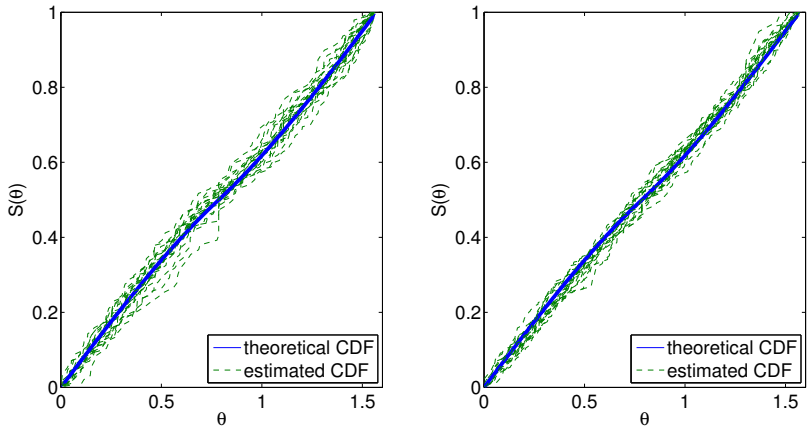

Figure 3.4: Estimated spectral measures from 20 simulations of Example 3.3.3 (a) with $n = 5000$ and $n = 20000$

n	$\alpha^{(1)}$	$\alpha^{(2)}$	$N^{(1)}$	$N^{(2)}$	$\bar{N}$
1000	1.1294	1.1283	141.9	147.5	101.3
5000	1.0734	1.0712	239.9	219.1	152.8
20000	1.0282	1.0261	303.4	318.5	204.1

Table 3.10: Simulation results for Example 3.3.3 (b)

n	Adaptive bias	Adaptive variance	5% bias	5% variance
1000	.0023	.0020	.0023	.0021
5000	.0012	.0013	.0025	$4.069 * 10^{-4}$
20000	$2.461 * 10^{-4}$	$8.861 * 10^{-4}$	.0026	$9.784 * 10^{-5}$

Table 3.11: Simulation results for Example 3.3.3(b)

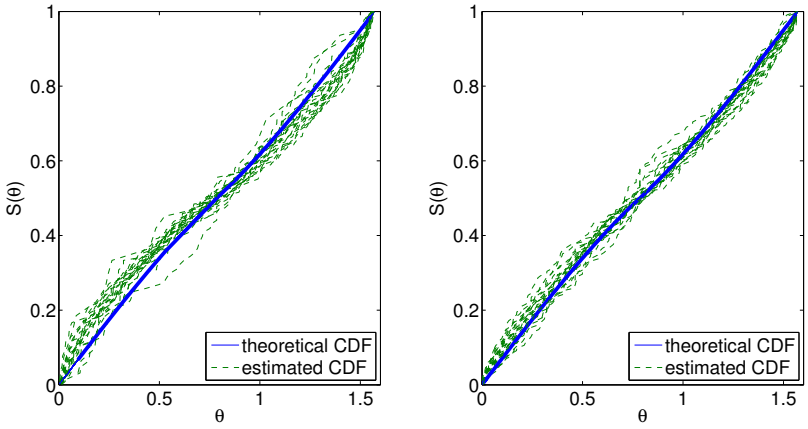

Figure 3.5: Estimated spectral measures from 20 simulations of Example 3.3.3 (b) with $n = 5000$ and $n = 20000$

5% rule works very well in all situations. This is because the entire distribution is rotationally invariant and, hence, in this special case all parts of the sample provide exactly the same information, and one does not need to detect the tail part of the sample. In contrast, in the "contaminated" case of part (b), it is important to know where the tails begin, and our approach, again, is superior to the 5% rule, apart from the smallest sample size situation.

It is sometimes of interest to estimate the actual density of the spectral measure, as opposed to the CDF. We use a kernel density estimation procedure with the values of $\theta_{i,N}$ in (3.5) such that $R_{i,N} > 1$ as the input. Since the spectral measure is concentrated on the interval $[0, \pi/2]$, we use the beta kernel of Chen [1999] (with bandwidth $b = .1$) that takes into account the proximity of an observation to the endpoints of the interval. The results of estimating the spectral density for the contaminated Cauchy example are presented in Figure 3.6. Since the density, as a derivative of the CDF, is more sensitive to the noise in the data than the latter, it is not surprising that only at relatively large sample sizes we obtain a reasonable fit. Even here the true spectral density is estimated much better by our procedure

than using the 5% rule.

3.4 Application to CoVaR

In this section we apply our approach of estimating multivariate tails to the problem of evaluating the conditional risk measure $CoVaR$. It is based on the widely used risk measure of Value-at-Risk.

Recall that for $0 < q < 1$, and close to 1, $VaR_q^{(j)}$ is defined as the q-quantile of the loss distribution $F^{(j)}$ of the jth asset. To be precise,

$$VaR_q^{(j)} = F^{(j)\leftarrow}(q)$$

(here and in the sequel we will assume, for simplicity of notation, that the appropriate distributions are continuous). In light of the recent financial crisis, efforts have been made to capture better the spillover effects between financial institutions. This gave rise to the conditional risk measure $CoVaR$ defined by

$$P\big(X^{(j)} \geq CoVaR_q^{j|i} \big| C(X^{(i)})\big) = 1 - q\,. \tag{3.9}$$

or

$$CoVaR_q^{j|i} = VaR_q\big(X^{(j)} \big| C(X^{(i)})\big)\,.$$

Here $C(X^{(i)})$ is some event describing a measure of distress of institution i. In this section we will consider events of the type $X^i \geq VaR_r^{(i)}$ for some $0 < r < 1$, close to 1. We are also interested in estimating the natural versions of $CoVaR$ in the case of more than one institution being in distress.

The original definition of CoVaR proposed by Adrian and Brunnermeier [2011] used conditioning on $X^i = VaR_r^{(i)}$. Improvements to this definition were suggested

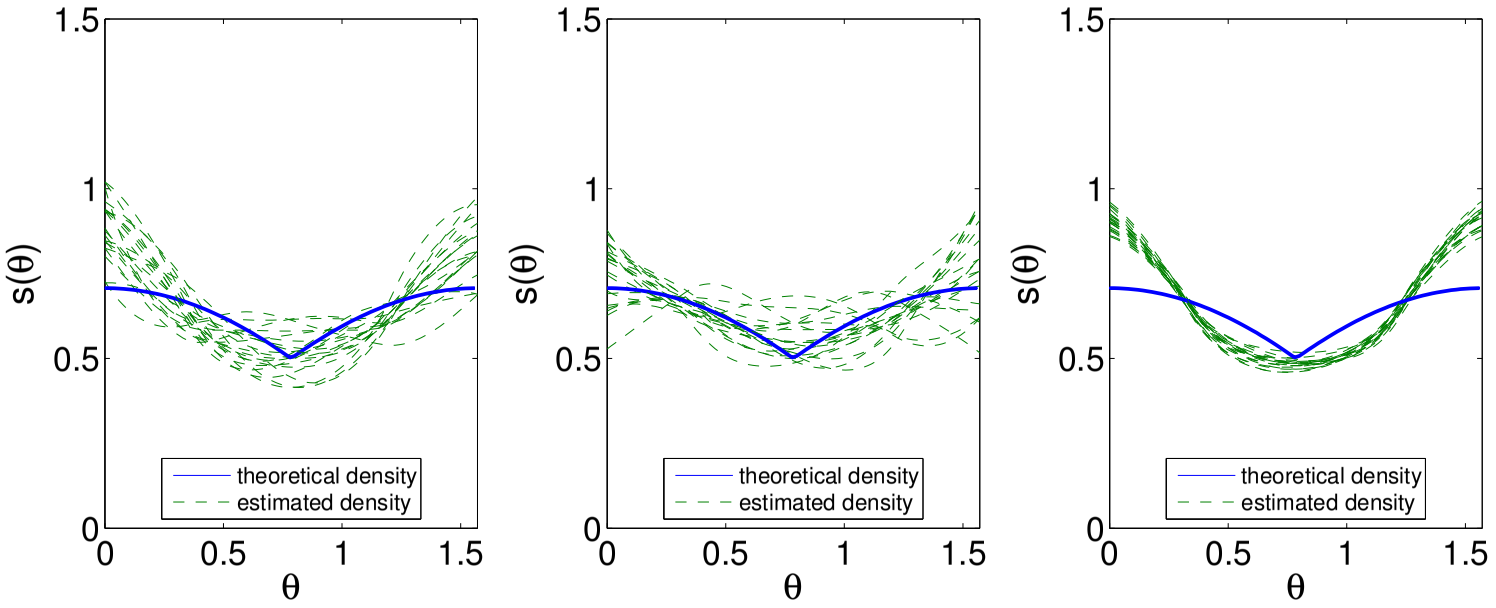

Figure 3.6: 20 kernel estimations for the density function of contaminated Cauchy tail with (from left to right) (1) n=5000 using our adaptive method, (2) n=20000 using our adaptive method, (3) n=20000 using the 5% rule

by Klyman [2011]. The inconsistency of the original definition was pointed out by Girardi and Ergun [2012]. The alternative version of CoVaR to involving conditioning on $X^i \geq VaR_r^{(i)}$ was later proved to be more consistent as a risk measure by Mainik and Schaanning [2012]. Other systemic risk measures were proposed by Zhou [2010]; the estimation procedure there also uses the multivariate extreme value theory.

We approximate the marginal quantiles $VaR_r^{(i)}$, $i = 1, \ldots, d$ using the Generalized Pareto model (3.6) of Example 3.3.1 (b). As before, we set the location parameter to be the $N_n^{(i)}$th order statistic in (3.3), and the remaining two parameters by the maximum likelihood estimation. Consequently, we estimate, using (1.6), the probabilities in (3.9) for t of the order $u_j(1/(1-r))$ (where u_j is the jth marginal generalized inverse function) as follows.

$$
\begin{aligned}
&P\big(Y^{(j)} > t \big| Y^{(i)} > VaR_r^{(i)},\, i = 1, \ldots, d,\, i \neq j\big) \qquad (3.10)\\
&= \frac{P\big(\frac{1-r}{(1-F^{(j)}(Y^j))} > \frac{1-r}{(1-F^{(j)}(t))}, \frac{1-r}{(1-F^{(i)}(Y^i))} > 1\,, i = 1, \ldots, d,\, i \neq j\big)}{P\big(\frac{1-r}{(1-F^{(i)}(Y^i))} > 1\,, i = 1, \ldots, d,\, i \neq j\big)}\\
&\approx \frac{\nu_*\big(y \in \mathbf{R}^d : y^{(j)} > \frac{1-r}{1-F^{(j)}(t)}, y^{(i)} > 1\,, i = 1, \ldots, d,\, i \neq j\big)}{\nu_*\big(y \in \mathbf{R}^d : y^{(i)} > 1, i = 1, \ldots, d,\, i \neq j\big)}\,.
\end{aligned}
$$

We mention that this approximation is not useful in the case of asymptotic independence, when the tail measure ν_* is concentrated on the axes; see e.g. Ledford and Tawn [1996]. In this case the approximation in (3.10) is not needed.

When using the estimator (3.10) in practice, one needs to deal with the following issue. Since we would like to estimate high conditional quantiles, we will need to use the estimator for t such that $(1-r)/\big(1-F^{(j)}(t)\big)$ is very large. In such cases the sample evaluation of the value of the tail measure ν_* in the numerator of the ratio in the right hand side of (3.10), described in Theorem 1, may be based

on a very small number of observations. We have chosen to resolve this difficulty by using the scaling property (1.8) of ν_*: for every $L > 0$ we can rewrite the ratio in the right hand side of (3.10) as

$$\frac{\nu_*\big(y \in \mathbf{R}^d : y^{(j)} > \dfrac{1-r}{1-F^{(j)}(t)}\dfrac{1}{L}, y^{(i)} > \dfrac{1}{L},\, i = 1, \ldots, d,\, i \neq j\big)}{L\nu_*\big(y \in \mathbf{R}^d : y^{(i)} > 1,\, i = 1, \ldots, d,\, i \neq j\big)}\,. \tag{3.11}$$

By choosing an appropriately large L, the difficulty described above is reduced. Another difficulty arises, however. Notice that an appropriate choice of L is, by the nature of the problem, t-dependent. We are estimating a function of t that is, clearly, monotone, and so is the expression in (3.11). However, we are using empirical estimates of ν_* both in the numerator and the denominator of (3.11). With a t-dependent L, the monotonicity may be occasionally violated. We have found that a choice of L, that seems to reduce the frequency and the effect of such violations, is

$$L_t = \left\| \left(\frac{1-r}{1-F^{(j)}(t)}, 1, \ldots, 1 \right) \right\|_2 .$$

We demonstrate our procedure on two examples, that show markedly different behavior of *CoVaR*: daily net returns of shares of 4 major US financial institutions and daily net returns of 4 major European indices.

Example 3.4.1. *Daily returns of 4 major US financial companies*

We analyze the daily returns on the stock of Bank of America, JP Morgan, Morgan Stanley and Citigroup, from January 2, 1996 to June 30, 2012. The data set contains 4169 observations.

As described above, we concentrate on the respective losses $Y^{(i)}$, $i = 1, \ldots, 4$ of these 4 stocks. We use as the distress events C_i the exceedance events $Y^{(i)} > VaR_r^{(i)}$, $i = 2, 3, 4$ with $r = .95$ and $r = .99$, and estimate the conditional

Quantile	Unconditional	Given C_2	Given C_2, C_3	Given C_2, C_3, C_4
90%	.026	.116	.157	.174
95%	.039	.157	.218	.238
99%	.090	.292	.380	.411

Table 3.12: Estimated quantiles for losses of Bank of America, unconditional and conditioned on $Y^{(i)} > VaR_{.95}(Y^{(i)})$.

quantiles of the losses on the stock of Bank of America. We specifically concentrate on the change in these conditional quantiles as we condition on more and more distress events for other companies. That is, we will estimate the conditional probabilities $P\big(Y^{(1)} > t|C_2\big)$, $P\big(Y^{(1)} > t|C_2, C_3\big)$ and $P\big(Y^{(1)} > t|C_2, C_3, C_4\big)$, as well as the resulting conditional quantiles. For comparison, we estimate unconditional quantiles as well.

The results are presented in Figure 3.7. Note that including every additional distress event adds mass to the conditional tail of the loss distribution on the stock of Bank of America and affects, correspondingly, the resulting value of *CoVaR*.

Due to the difference in the strength of dependence among the 4 assets, permutations of the order of $Y^{(2)}, Y^{(3)}, Y^{(4)}$ in the conditioning would shift the middle curves in Figure 3.7 slightly to the right or left

Some estimated quantiles for losses of Bank of America are reported in Table 3.12.

Example 3.4.2. *Daily returns of 4 major European indices*

We analyze the daily returns of the following indices: British FTSE 100, French CAC 40, Deutsche Börse AG and Spain's IBEX 35, from January 2, 1995 to April 27, 2012. The data set contains 4402 observations.

Once again, we concentrate on the respective losses $Y^{(i)} = -X^{(i)}, i = 1, \ldots, 4$.

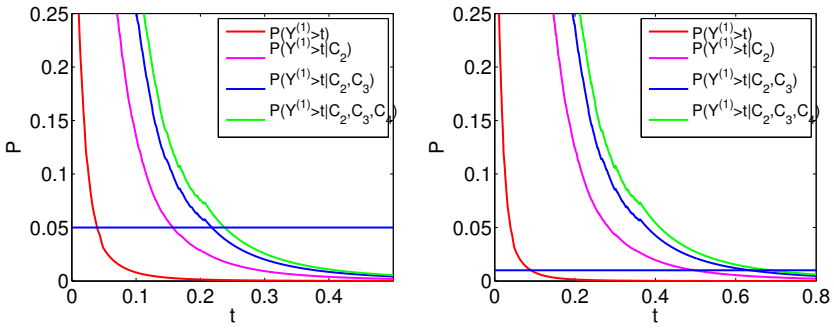

Figure 3.7: US stocks: Unconditional $P(Y^{(1)} > t)$ and conditional $P(Y^{(1)} > t|C_i)$ with $C_i = \{Y^{(i)} > VaR_r(Y^{(i)})\}$ for $r = .95$ and $r = .99$

Quantile	Unconditional	Cond. on $Y^{(2)}$	on $Y^{(2)}, Y^{(3)}$	on $Y^{(2)}, Y^{(3)}, Y^{(4)}$
90%	.013	.043	.049	.050
95%	.019	.053	.059	.060
99%	.035	.078	.086	.087

Table 3.13: Estimated quantiles for losses of FTSE 100 unconditioned and conditioned on $Y^{(i)} > VaR_{.95}(Y^{(i)})$

Using the same type of distress events as in the previous examples, we study the conditional loss tail of the return on British FTSE 100 and the corresponding *CoVaR*. The results are displayed in Figure 3.8 and Table 3.13. An obvious difference between the results obtained in this example and the previous example is that no significant movement in the conditional loss probabilities is noticeable after the first conditioning. This is due to the fact that there is strong tail dependence between the losses on French CAC 40, Deutsche Börse AG and Spain's IBEX 35. This results in the tail measure that concentrated in a small part of the four-dimensional space, and in the conditional loss curves that barely move.

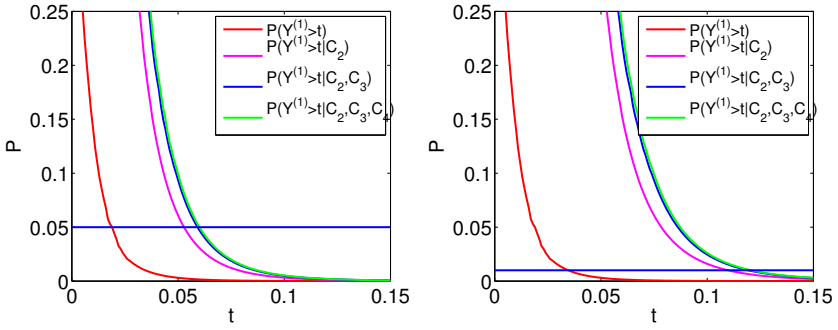

Figure 3.8: European indices: Unconditional and conditional $P(Y^{(1)} > t|C_i)$ where C_i is the event $Y^{(i)} > VaR_r(Y^{(i)})$ for $r = .95$ and $r = .99$

CHAPTER 4

ESTIMATING SPECTRAL MEASURE WITH HIGH DIMENSIONAL DATA

4.1 The curse of dimensionality

The "curse of dimensionality" is a phrase coined by Richard Bellman (Bellman et al. [1966]) in the context of optimization by exhaustive enumeration. The phrase mainly refers to the fact that if we consider a Cartesian grid of 1/10 length on a unit interval, we have 10 points. If we consider this grid on a 2 dimensional unit square, we have 100 points and for 10 dimensional unit cube, we would have 10^{10} points. So to crudely optimize a continuous function on the unit cube by estimating the function at different points on this grid would soon lead us to an insurmountable number of calculations.

It is easily seen that this "curse of dimensionality" applies not only in optimization but also in statistical estimation. Assuming that we need at least 10 data points to get a decent density estimation on the unit interval, estimating probabilities on a d-dimensional unit cube would require us to have at least 10^d data points, which in many cases is not feasible. In the context of tail measure, we seem to run into this problem of lack of data twice as the majority of the data would be consider redundant and only a small fraction of tail data is used in the estimation.

The challenge that comes with high dimensional data in statistical estimation is not only in the form of lack of data. Nonparametric density estimators like the kernel density estimator only work well up to a dimension of $d = 3$. With

an appropriate bandwidth selection, the optimal rate of convergence of MSE of a kernel density estimation to 0 is $\mathcal{O}(n^{-4/(d+4)})$ ([Wand and Jones, 1994b]). This gives us a very slow rate of convergence when the dimension is high. This is not to mention the high computational cost associated with finding the optimal bandwidth selection like smooth cross validation or plug-in selectors ([Wand and Jones, 1994a], [Duong and Hazelton, 2005], [Hall et al., 1992]).

Parametric modeling using the extreme value distribution, in the context of component wise maxima, has been studied by [Tawn, 1988] for bivariate data and by [Stephenson and Tawn, 2005] for bivariate and trivariate data with a focus on logistic models. [Stephenson, 2009] proposed a method to deal with higher dimensions using asymmetric logistic model that can be applied to both component wise maxima and exceedances over threshold.

There are some alternative approaches proposed to study multivariate extremes, but again, none of these methods can mitigate the curse of dimensionality. One approach is to fit a multivariate t distribution to the data ([Glasserman, 2003]). But just like the univariate version, in exchange for its simple formula, t- distribution has many inflexibilities and may not be the best fit for the data. Another approach involves a concept called "copula". Copula is simply a joint distribution function on the unit cube that captures the dependence structure of the random variables. Suppose $(X_1, X_2, ..., X_d)$ is a random vector and F_i, $i = 1, ..., d$ are continuous marginal distributions. Then the copula C is defined as follows

$$C(u_1, u_2, ..., u_d) = P(X_1 \leq F_1^{\leftarrow}(u_1), X_2 \leq F_2^{\leftarrow}(u_2), ..., X_d \leq F_d^{\leftarrow}(u_d))$$

$$\text{for } 0 \leq u_1, u_2, ..., u_d \leq 1$$

There is a tremendous amount of research on this topic as copulas are used widely in financial industry from risk and portfolio management to pricing derivatives. .

For surveys on copulas and its application in finance, see [Bouyé et al., 2000], [Embrechts et al., 2002],[Cherubini et al., 2004], [Trivedi and Zimmer, 2007]. There are several criticisms on the use of copula, especially the Gaussian copula, which was believed to lead to widespread risk modeling troubles in the industry ([Whitehouse, 2005]). A discussion on the validity of copula models can be seen in [Mikosch, 2006], followed by a rejoinder.

4.2 Sparsity of financial spectral measure

Financial spectral measure, in our observation, does not distribute evenly on the unit sphere. Nor do the extreme data seem to cluster in a few low dimensional subspaces which would enable us to use current clustering or dimension reduction techniques to study them.

To study the spectral measure of financial data in the high dimensional, we use the rank transform method described in Chapter 3. For the jth marginal, $j = 1, \ldots, d$ we calculate the first rejection time, $N^{(j)}$, via

$$N_n^{(j)} := \inf \left\{ k : 1 \leq k \leq n,\, |Q_{k,n}| \geq \omega_j \sqrt{\frac{\lambda_n^{(j)}}{k}} \right\} ,$$

and set

$$N_n = \bigwedge_{j=1}^{d} N_n^{(j)} .$$

We then "rank transform" the original data $Z_i^{(j)}, i = 1, ..., n, j = 1, ..., d$ into $\dfrac{N_n}{r_i^{(j)}}$ where

$$r_i^j = \sum_{m=1}^{n} 1_{[Z_m^{(j)} \geq Z_i^{(j)}]}, \quad 1 \leq i \leq n ,$$

The extreme data in which we will base our analysis on are those transformed points with radius greater or equal to 1. In this chapter, we will still be using

l^∞, hence data points are considered extreme if any one of their coordinates are greater or equal to 1. After this step, we rescale the extreme points into unit norm and these serve as the basis of our spectral measure analysis.

In theory, on the unit cube, we say, for example, that a point belongs to the 0-dimensional space form by the cube and an axis if it has exactly one coordinate of unity and the rest are zeros. Similarly, a point belong to a 1-dimensional space on the cube created by 2 axes have exactly 2 nonzero coordinates. So the number of nonzero coordinates in each extreme point is in effect a description of the number of dimension of the subspace that point belong to. For real life data, we allow a small margin of error and only count the number of significant coordinates of each extreme point. By "significant", we mean any coordinate that is greater than a small ϵ, say $\epsilon = .1$. This choice of $\epsilon = .1$ is mostly arbitrary, but it is conservative enough that for an extreme point to be deemed belonging on an axis, the largest coordinate has to be at least 10 times larger than all of the other coordinates.

We then plot a histogram of the number significant coordinates contained in each data point.

There are two major characteristics of extreme financial data that we observe. First, most of the extremes lie in low dimensional subspaces; at the very least, half of the tail data would belong to a subspace that has at most 3 significant coordinates. In fact, a significant percentage of extreme data lies on the axes. Second, the extreme data lies in a large number of subspaces, but each subspace contains very few data points.

We begin with a dataset of absolute daily return for SP 500 stocks, spanning a period from January 1, 1996 to December 31, 2003. The dataset contains approx-

imately 2000 data points.

Figure 4.1 shows such a histogram for 5 stocks from the index: Comverse Technology, Dell, Coca Cola, Consolidated Edison and Exxon Mobil. The extreme data consists of 274 points. As we described earlier, most of the extremes lie in a subspace with less than 3 significant coordinates. There are in total, 106 extreme points with only one significant coordinate. These points are the points clustering around the axis with each axis ranging between 19 to 25 points. There are 95 extreme points with 2 significant coordinates. If we number the stocks from 1 to 5 respectively, we can tally these counts in Table 4.1.

We can see that the 5 stocks can be split into 2 groups, the first two are technology companies that move together in the extreme. The other three, large companies producing drink, electricity and energy respectively also move together. There is some cross dependency between the two groups, especially between Comverse and stocks from group 2, but not to the same extent as within group. Sets of points with 4 significant coordinates have 2-4 counts each. This supports our description of the support of extreme data as a union of a large number of smaller dimension subspaces but each with very few data points.

Figure 4.2 shows a histogram for 5 finance stocks from the index: Morgan Stanley, Bank of New York Mellon, American Express, Capital One and Zions Bancorporation. The slower rate of decay in the histogram reflects a higher dependency in the extreme among the stocks. Nevertheless, data with no more than 3 significant coordinates account for approximately 70% of the extreme data.

We now study another data set of absolute log minute returns of stocks from SP500 over a period of 754 trading days, from January 2007 to December 2009.

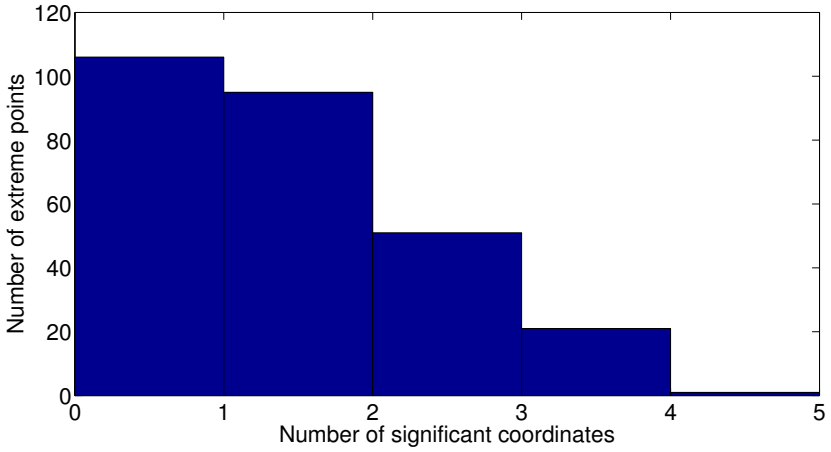

Figure 4.1: Histogram of number of significant coordinates in tail data of 5 stocks (daily returns)

Significant coordinates	Count	Significant coordinates	Count
1,2	18	1,2,3	2
1,3	12	1,2,4	5
1,4	7	1,2,5	3
1,5	11	1,3,4	3
2,3	9	1,3,5	5
2,4	6	1,4,5	4
2,5	2	2,3,4	5
3,4	10	2,3,5	5
3,5	11	2,4,5	5
4,5	9	3,4,5	14

Table 4.1: Counts of number of significant coordinates in extreme data of 5 stocks (daily returns)

The data set contains 465 stocks and approximately 288000 data points.

We produce histograms of the number of significant coordinates for sets of the first 10, 50 and 100 stocks by alphabetical order from the dataset. These are shown in Figure 4.3, Figure 4.4 and Figure 4.5 respectively.

For Figure 4.3, the 10 stocks used are: 3M, Abott Laboratories, Abercrombie & Fitch, ACE Limited, Adobe, AES Corporation, AFLAC, Agilent Technology and AK Steel. There are around 50 data points clustering around each axis. The counts

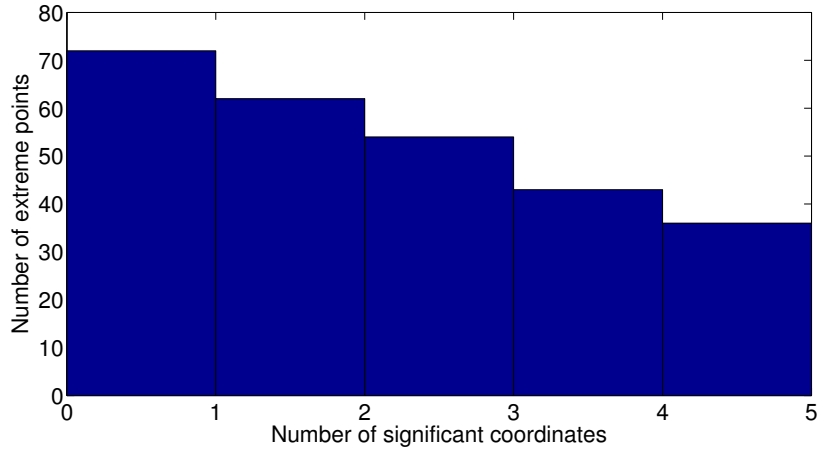

Figure 4.2: Histogram of number of significant coordinates in tail data of 5 banking stocks (daily returns)

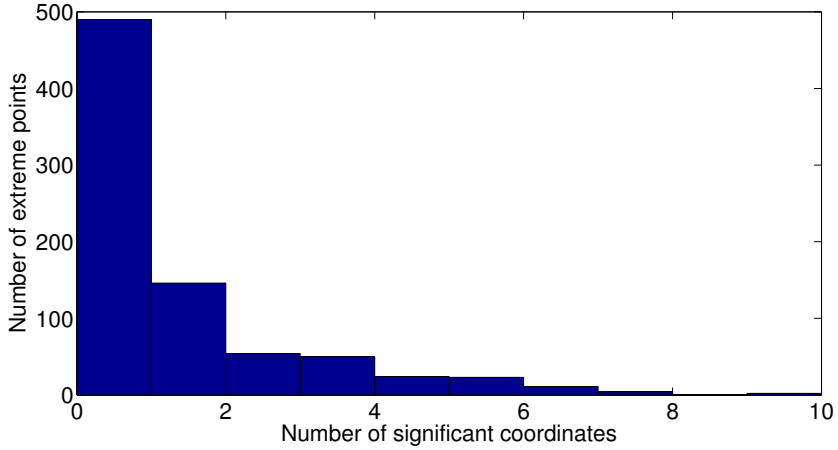

Figure 4.3: Histogram of number of significant coordinates in 10 dimensional space (minute returns)

drop significantly for those with 2 significant coordinates. For 38 combinations of 2 stocks out of the possible 45, the count for each combination range between 1 and 4; the other 7 possible combinations have a maximum of 8 data points. For combinations of higher number of stocks, the majority of possible combinations have 0 observation while the others have a maximum of 3 observations. The sparsity is a lot more pronounced than daily returns data. This is expected since extreme comovements in the same minute by different stocks should occur with

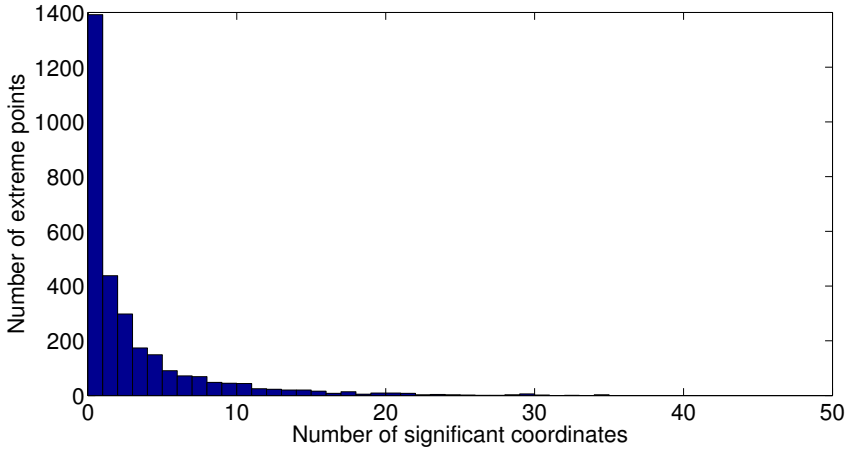

Figure 4.4: Histogram of number of significant coordinates in 50 dimensional space (minute returns)

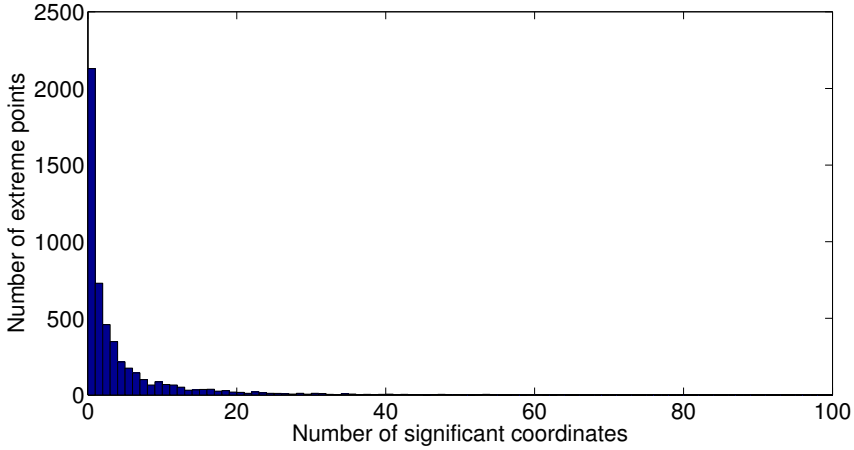

Figure 4.5: Histogram of number of significant coordinates in 100 dimensional space (minute returns)

lower frequency than those in the same day.

One can see from Figure 4.3, Figure 4.4 and Figure 4.5 that a large percentage of extreme data have only 1 significant coordinate. In fact, data with at most 3 significant coordinates comprise at least 80% of tail data. Many subspaces contain no data points.

The number of extreme data points in the 10-dimensional, 50-dimensional, 100-

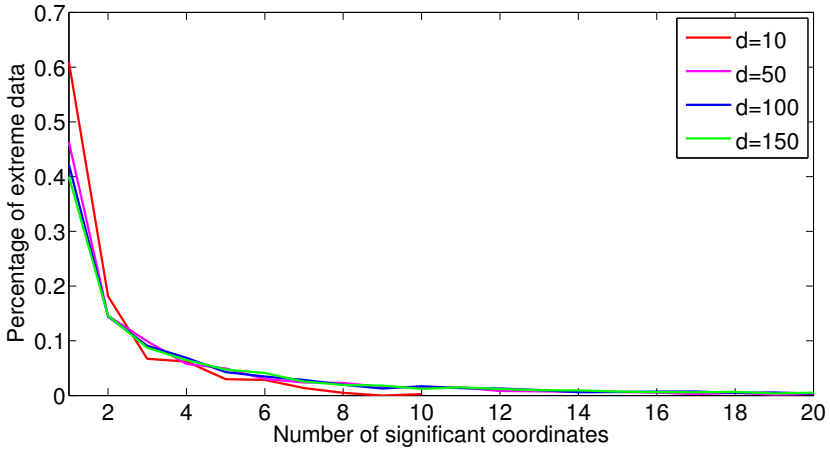

Figure 4.6: Histograms (minute returns) scaled and superimposed

dimensional cases are 804, 3006 and 5049 respectively. The more dimensions we include in our analysis, the more data points become a part of the extremes. This can be explained as the more stocks we have, the chances that any particular data point would contain at least an extreme coordinate of any of those stocks increases.

There is also a phenomenon of self similarity that we can observe from the structure of the tail. We rescale the histograms of Figure 4.3, Figure 4.4, Figure 4.5 and also one for the case of 150 dimensions so that the area underneath is unity. This rescaled plot is shown in Figure 4.6. The last three curves are nearly identical.

Our choice of $\lambda_n = \log n^2$ in the previous section is a recommendation for small and intermediate sample size. As in Theorem 2, the marginal stopping point k is roughly proportional to λ_n, and in the case of our very large sample size, $\lambda_n = \log n^2 \approx 150$. This constitutes as less than .05% of the data. We will now try to relax the choice of λ_n to $\log n^3 \approx 2000$. As we relax the condition of what the tail entails, more data in the middle is considered extreme, hence there is a larger degree of dependence among the stocks, shown in the slightly slower rate

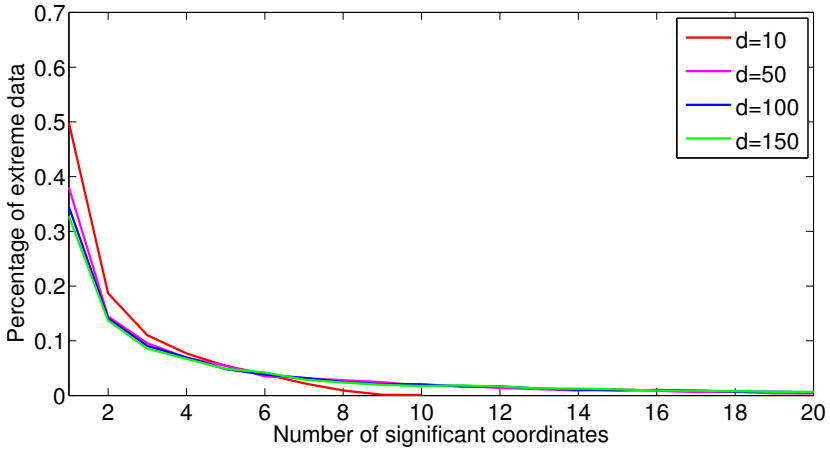

Figure 4.7: Histogram (minute returns) scaled and superimposed with $\lambda_n = \log(n)^3$

of decay of the rescaled histograms in Figure 4.7. Nevertheless, the self similarity phenomenon is observed. This seems to suggest that as long as we keep λ_n the same in all dimensions, the more dimensions we add to our analysis, the percentage of extreme data that lies in lower dimensional subspaces stay unchanged.

We also include in Figure 4.8 the histogram for 10 finance stocks and in Figure 4.9 histogram for 10 health stocks. As we expected, the dependency among stocks is higher than that shown in Figure 4.3 but not significantly so and nowhere near the level of dependency that we observed for daily returns. Nevertheless, the two observations we have about the tail structure, that the majority of the extremes are low dimensional and that they belong to a large number of subspaces, each with very few observations, still hold true.

Remark 7. *One drawback of our procedure which we discovered during this analysis is that it is extremely sensitive to corrupted data. Since our sequential testing relies on the largest order statistics, then if there is an error in the data which results in an extreme outlier compared to the rest of the data this outlier can cause the test to terminate early.*

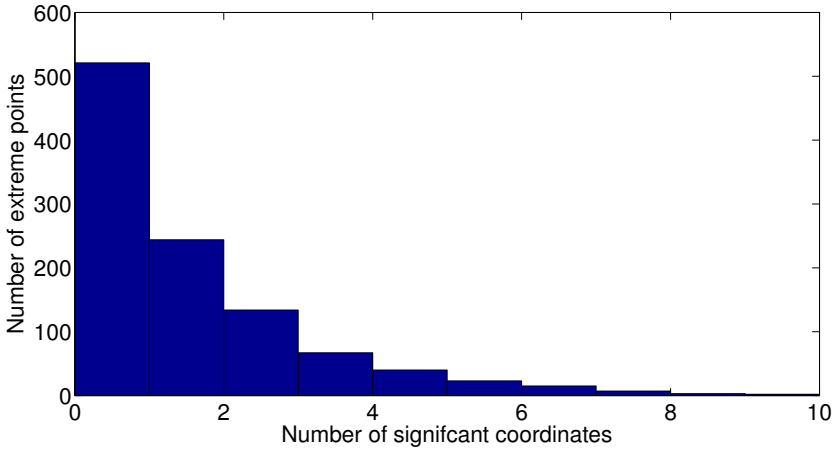

Figure 4.8: Histogram of number of significant coordinates in tail data of 10 finance stocks (minute returns)

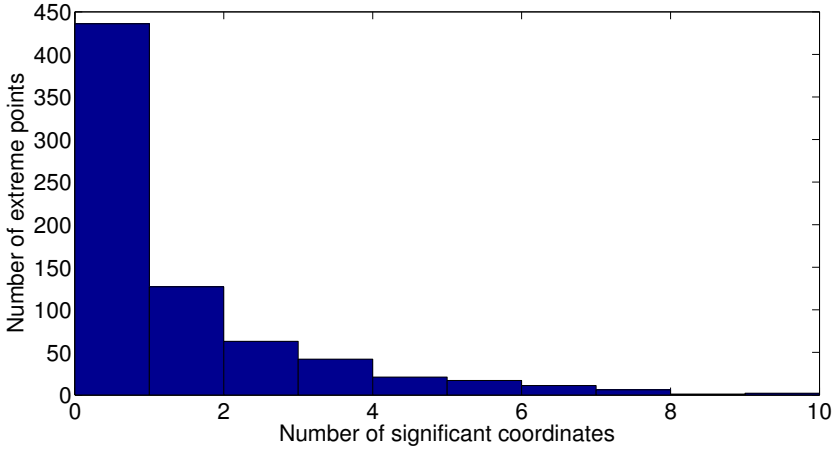

Figure 4.9: Histogram of number of significant coordinate in tail data of 10 health stocks (minute returns)

An example of this drawback is the returns from Aetna. In our dataset, a sequence of 3 consecutive minute prices was recorded as follow: 45.93, .20, 45.85. This single error caused our sequential testing with $\lambda_n = \log n^2 \approx 150$ *to stop at* $k = 16$*. If this entry is removed, the sequential test would instead give* $k = 451$*, a marked difference. This phenomenon is not isolated, there are other stocks in our data set where a single faulty entry results in an outsized return which then "contaminates" our tests.*

4.3 Estimating spectral measures in high dimensions

As is evidenced in the previous section, the spectral measure in financial data sets is supported, in a large part, on a large number of low dimensional subspaces. Hence in a given sample, we generally have very few, or none, of extreme observations in most subspaces.

How should we proceed with such a data set? Our goal is to construct a nonparametric model in which we work with subspaces of low dimension (e.g., no more than 3), because only with such a model can we hope to arrive at some reasonable estimations. This is not an unreasonable target as data with no more than 4 significant components comprises a large part of the extreme data set.

The approach we suggest is to break the data set into a union of smaller dimensional subspaces and apply a mixture model. In effect, our high dimensional spectral measure would be treated as a weighted sum of lower dimensional spectral measures. This approach has the unfortunate effect of causing certain discontinuities in the spectral measure when we move from one subspace to another. On the other hand, the advantages it brings are significant. First, modeling in low dimensional subspaces can be easily done with existing statistical tools. Second, by breaking up the spectral measure into several smaller subspaces, one can look at the estimation and have an intuitive understanding of which assets, or combinations of assets, are driving the extreme movements.

We begin with the first step of specifying the probability weight of each subspace. Next, we estimate the spectral measure within that subspace - for those up to dimension 3, this is easily achieved with standard density estimation. We do not profess to being able to completely overcome the curse of dimensionality

with this approach. We are still not able to estimate the probability density for the small part of the data that lies in very high dimensional subspaces. But at the very least, we can assign a probability weight to such subspaces. This would give us a rough idea of whether or not such combination of assets may often have large movements at the same time. We will now focus on the first step and propose a method to accomplish it.

Our idea for assigning probability weights stems from the fact that due to the lack and also sparsity of the data, it is not unusual that we see subspaces with very few observations. But for all intents and purposes, with the lack of extreme data, there is very little difference, for example, between a 4-dimensional subspace with 0 observation and one with 1 observation. We believe everything happens in proportion and want to give the subspaces with very few observations a slight equalizer effect with the following logic. Suppose we have a data point in which X_1, X_2, X_3 are all significant, and X_1, X_2 are the largest coordinates. It is not farfetched that we may encounter significant movements in the lower dimensional subspaces (X_1) or (X_1, X_2). In other words, when we calculate probability weights, we not only take in to account the observed data point but also their largest coordinates. This idea will be explained more formally as follows.

Suppose we want to calculate the weight of the interior of the subspace $X_{i_1}, ..., X_{i_k}$. By interior, we mean that it is the subspace spanned by $X_{i_1}, ..., X_{i_k}$ excluding all smaller dimension subspaces spanned by any subsets of the set described. We say a coordinate is significant if its magnitude is greater or equal to ϵ, which we set to be .1. Let $n_{i_1,...,i_k}$ be the number of tail data points in which $X_{i_1}, ..., X_{i_k}$ are the k largest significant coordinates, n be the number of extreme data points. We also define n_k to be the number of extreme data points that has

exactly k significant coordinates and $n_{(k)}$ to be the number of extreme data points that has at least k significant coordinates. We then set the weight of the interior of subspace $X_{i_1}, ..., X_{i_k}$ as follows

$$w_{i_1,...,i_k} = \frac{n_{i_1,...,i_k}}{n_{(k)}} \cdot \frac{n_k}{n} \tag{4.1}$$

One can easily see that the sum of weight of all subspace interiors adds up to 1. Suppose A_k is the set of all possible combinations of k assets. Then we have

$$\begin{aligned}
\sum_{k=1}^{d} \sum_{A_k} w_{i_1,...,i_k} &= \sum_{k=1}^{d} \sum_{A_k} \frac{n_{i_1,...,i_k}}{n_{(k)}} \cdot \frac{n_k}{n} \\
&= \sum_{k=1}^{d} \frac{n_k}{n} \left(\sum_{A_k} \frac{n_{i_1,...,i_k}}{n_{(k)}} \right) \\
&= \sum_{k=1}^{d} \frac{n_k}{n} \cdot 1 \\
&= 1
\end{aligned}$$

There are two different ways we can explain the basis of this formula. The first explanation comes from an empirical perspective. Notice that in this weight formula is just a slight modification away from the standard empirical estimation

$$w^{\text{standard}}_{i_1,...,i_k} = \frac{n^{\text{standard}}_{i_1,...,i_k}}{n} = \frac{n^{\text{standard}}_{i_1,...,i_k}}{n_k} \cdot \frac{n_k}{n} \tag{4.2}$$

where $n^{\text{standard}}_{i_1,...,i_k}$ is the number of data points with exactly k significant coordinates and those k coordinates are $i_1, i_2, ..., i_k$.

The difference between the 2 formulas lies in the first fraction of the product. For the standard empirical formula, this fraction estimates the probability that

given a data point is in a $k-1$ dimensional subspace, what is the probability that it lies in the specific subspace spanned by $i_1, i_2, ..., i_k$. In effect, the modified formula operates on the same principle, it is trying to calculate the same probability. The only difference is that it treats the data set as if we have more data than we actually do. Supposed we have observed data point with more than k significant coordinates and the largest one are $i_1, i_2, ..., i_k$, then our model behaves as if we have another observed point in which the only k significant coordinates are $i_1, i_2, ..., i_k$.

Another way to explain our formula is that we are making extra assumptions about the structure of the spectral measure. Let P_1 be the probability that the largest k coordinates are of a specific set given a random variable X has exactly k significant ones. Let P_2 be the probability that the largest k coordinates are of a specific set given X has at least k significant ones. We are making the assumption that P_1 and P_2 are the same. In order words, knowing that X has at least k significant coordinates is not different from knowing that it has exactly k significant coordinates with respect to determining what those coordinates are. With this assumption,

$$\begin{aligned} P(X \text{ is in interior subspace } \{i_1, ..., i_k\}) &= P_1 * P(X \text{ has } k \text{ significant coordinates}) \\ &= P_2 * P(X \text{ has } k \text{ significant coordinates}) \\ &\approx \frac{n_{i_1,...,i_k}}{n_{(k)}} \cdot \frac{n_k}{n} \end{aligned}$$

We use the daily returns of the five stocks from the previous section (Comverse Technology, Dell, Coca Cola, Consolidated Edison and Exxon Mobil) to demonstrate the effect of our formula compared to the standard empirical distribution. The result is shown in Table 4.2 where for each combination of $i_1, ..., i_k$, we report the exact number of data points observed in the interior of such subspace as well as $n_{i_1,...,i_k}$. We also report the standard empirical probability weights and the

$i_1, ..., i_k$	Exact number of data	$n_{i_1,...,i_k}$	Empirical weight	Modified weight
1	20	57	.0729	.0804
2	25	55	.0912	.0777
3	21	55	.0766	.0777
4	19	49	.0693	.0692
5	21	58	.0766	.0820
1,2	18	26	.0657	.0537
1,3	12	17	.0438	.0351
1,4	7	14	.0255	.0289
1,5	11	16	.0401	.0330
2,3	9	16	.0328	.0330
2,4	6	13	.0219	.0268
2,5	2	9	.0073	.0186
3,4	10	19	.0365	.0392
3,5	11	20	.0401	.0413
4,5	9	18	.0328	.0371
1,2,3	2	3	.0073	.0076
1,2,4	5	9	.0182	.0229
1,2,5	3	6	.0109	.0153
1,3,4	3	5	.0109	.0127
1,3,5	5	7	.0182	.0178
1,4,5	4	6	.0146	.0153
2,3,4	5	8	.0182	.0204
2,3,5	5	7	.0183	.0178
2,4,5	5	7	.0183	.0178
3,4,5	14	15	.0511	.0382
1,2,3,4	4	4	.0146	.0139
1,2,3,5	4	4	.0146	.0139
1,2,4,5	7	7	.0255	.0244
1,3,4,5	4	4	.0146	.0139
2,3,4,5	2	3	.0073	.0104
1,2,3,4,5	1	1	.0036	.0036

Table 4.2: Empirical and modified weights for 5 stocks (daily returns) with $\epsilon = .1$.

modified weights of Equation 4.1.

The equalizer effect is seen most clearly in the lower dimensional subspaces. For example, the subspace spanned by stock number 3 (Coca Cola) has only 21 observations compared to 25 of Dell, stock number 2. But, Coca Cola happens in

$i_1, ..., i_k$	Exact number of data	$n_{i_1,...,i_k}$	Empirical weight	Modified weight
6	6	27	.0162	.0177
7	9	39	.0243	.0255
5,10	0	3	0	.0024
6,7	1	6	.0027	.0047
1,3,6	0	1	0	.0011
1,2,7,10	2	4	.0054	.0046
3,4,5,6,7,8	0	1	0	.0017
2,3,4,5,6,7,8,9	1	1	.0269	.0269

Table 4.3: Empirical and modified weights for 10 stocks (daily returns).

higher frequency to be the one with the most extreme movement when it moved with other stocks. This was taken into account and the subspace of Coca Cola is assigned an equal weight to that of Dell with the modified formula. By design, this effect only happens among subspace interiors with the same dimensions. The two methods yield the same result for the sum of weights of all subspace interiors of dimension 0, or sum of weights of all subspace interiors of dimension 1 and so on.

For this example, there is no subspace with 0 observation, but one can easily see that with higher dimensional datasets, where the chance of 0-observational subspaces is high, some of those subspaces would be assign a minuscule weight even though there is no observation. We show an example of this by adding another 5 stocks to the above: General Electric, Novellus Systems, General Motor, Goodrich Corp. and Ingersoll-Rand. In Table 4.3, we display a sample of probability weights of some subspace interiors.

Once we calculate the weight of each subspace interiors, the remaining task to calculate density for each interior is simple. For example, the weight of the interior of Coca Cola and Dell subspace has a weight of .330. Our task is now to estimate a bivariate spectral measure which we know how to do. We come back to

the original dataset, discard all other dimensions and proceed as usual. Since our estimation is for the interior, once we perform the rank transform and unitizing the extremes, discard all data points that only have one significant coordinate and use the rest of the data for density estimation. Figure 4.10 show this interior spectral measure estimation using the beta kernel estimator with bandwidth .05. Here, θ is the angle from the polar transformation of Dell and Coca Cola. We see the dips of the density once the angle is close to 0 or $\pi/2$ since those data points so close to the axes are counted toward the 0-dimensional subspaces.

There is a reason to redo all the steps necessary to estimate the local spectral measure of Dell and Coca Cola instead of applying the kernel density estimation straight to the observed data we already identified in the weight assignment step. Recall that in this example, we have $N_n = \bigwedge_{j=1}^{5} N_n^{(j)}$. We also have $\bigwedge_{j=2}^{3} N_n^{(j)} \geq N_n$. We interpret this as in order to estimate the local spectral measure between X_2 and X_3, one does not need to set as high a radius threshold as when we try to estimate the global spectral measure of all X_i. We take advantage of this fact so that we can use more data from the lower threshold to estimate the density of the local spectral measure. In the case of Dell and Coca Cola, we arrive at 29 data points using the lower threshold, which is more than what we have for this subspace interior in the probability weight assignment step.

What happens if we are very conservative with ϵ and make it a lot smaller? We perform the same analysis of the five stocks (Comverse Technology, Dell, Coca Cola, Consolidated Edison and Exxon Mobil) but with $\epsilon = .05$. The assigned weights are shown in Table 4.4. The results are changed significantly as more points are now considered to belong in the higher dimensional spaces and weights are divided accordingly. Even so, 0,1 and 2-dimensional spaces make up 72% of

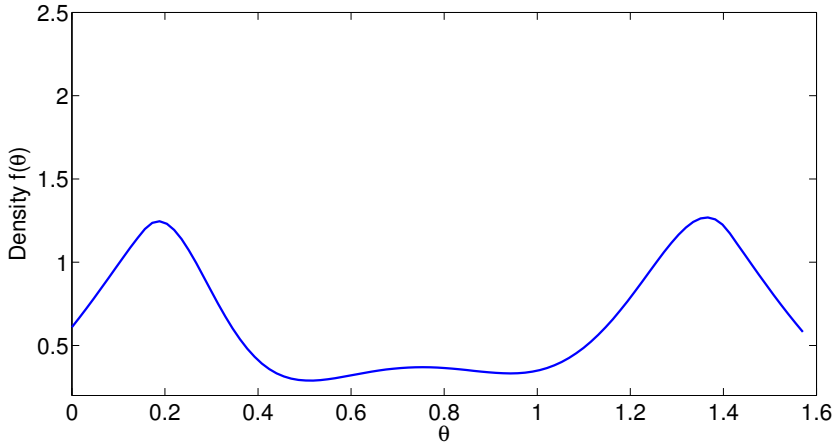

Figure 4.10: Interior spectral measure of Dell and Coca Cola with $\epsilon = .1$

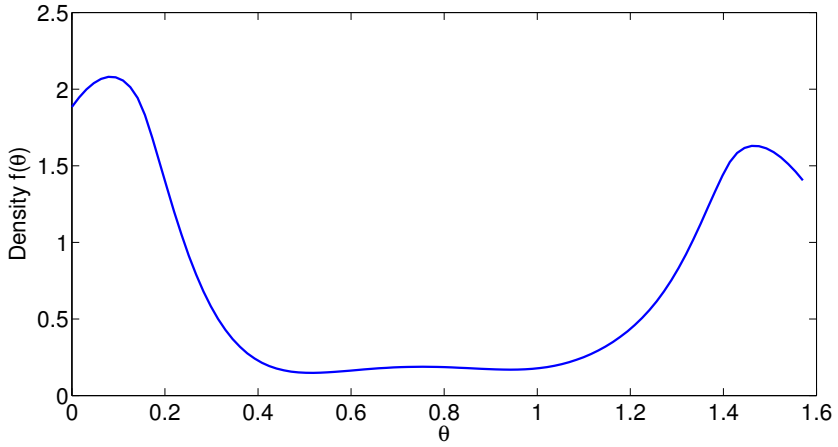

Figure 4.11: Interior spectral measure of Dell and Coca Cola with $\epsilon = .05$

all probability weights. The new spectral density of Dell and Coca Cola is shown in Figure 4.11. With the new ϵ, the interior of this subspace now has more than 50 data points used in the density estimation compared to 29 points before. All of the extra points lie near the axes hence the density becomes significantly larger when θ is near 0 or $\pi/2$.

One can easily deduce that if instead of making ϵ smaller, we enlarge it, we would get a spectral measure in which a very significant percentage of the probability weights fall on low dimensional subspaces. Local spectral measures would

$i_1, ..., i_k$	Exact number of data	$n_{i_1,...,i_k}$	Empirical weight	Modified weight
1	11	57	.0401	.0410
2	14	55	.0511	.0396
3	12	55	.0438	.0396
4	9	49	.0328	.0352
5	8	58	.0292	.0417
1,2	12	33	.0438	.0361
1,3	5	20	.0182	.0219
1,4	8	19	.0292	.0208
1,5	6	22	.0219	.0241
2,3	3	20	.0109	.0219
2,4	4	17	.0146	.0186
2,5	4	14	.0146	.0153
3,4	5	24	.0182	.0263
3,5	8	25	.0292	.0274
4,5	11	26	.0401	.0285
1,2,3	11	20	.0401	.0379
1,2,4	5	13	.0182	.0246
1,2,5	5	16	.0182	.0303
1,3,4	8	16	.0292	.0303
1,3,5	9	14	.0328	.0265
1,4,5	5	13	.0182	.0246
2,3,4	7	13	.0255	.0246
2,3,5	7	12	.0255	.0228
2,4,5	8	12	.0292	.0228
3,4,5	15	25	.0547	.0474
1,2,3,4	13	15	.0474	.0400
1,2,3,5	8	11	.0292	.0293
1,2,4,5	12	17	.0438	.0453
1,3,4,5	13	19	.0404	.0506
2,3,4,5	8	12	.0292	.0320
1,2,3,4,5	20	20	.0730	.0730

Table 4.4: Empirical and modified weights for 5 stocks (daily returns) with $\epsilon = .05$.

also have a lot less mass near the axes.

As a final note, we want to state again that this method is not exact. The probability weights fluctuates with how we define what a significant component entails. The method also does not tell us how to estimate probability density for the small part of the data that has a high number of significant coordinates. What the method does, however, is to provide us with a simple and intuitive way to understand the structure of a large portion of financial extreme data in high dimensions, which originally seemed to be a insurmountable task.

APPENDIX A

APPENDIX

A.1 Proof of Lemma 2.2.1

Proof. Fix $0 < \epsilon < |\rho|$. Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. Pareto(1) random variables (i.e. $P(Y_1 > y) = 1/y$ for $y \geq 1$). As in Eq. 3.2.7 in [de Haan and Ferreira, 2006], since $X_{n-i,n} \stackrel{d}{=} U(Y_{n-i,n})$, there exist function $A_0 \sim A$ and $r_0 > 0$ with the property that

$$\begin{aligned}
&\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}) + A_0(Y_{n-k,n})\frac{1}{\rho}\left(\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right)^{\rho} - 1\right) \\
&-\epsilon|A_0(Y_{n-k,n})|\frac{1}{\rho}\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right)^{\rho+\epsilon} \\
\leq &\frac{1}{\gamma}\log\frac{X_{n-i,n}}{X_{n-k,n}} \leq \\
&\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}) + A_0(Y_{n-k,n})\frac{1}{\rho}\left(\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right)^{\rho} - 1\right) \\
&+\epsilon|A_0(Y_{n-k,n})|\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right)^{\rho+\epsilon}
\end{aligned} \tag{A.1}$$

if $Y_{n-k,n} > r_0$ (this condition ensures that the observation is large enough for the second order regular variation condition to start playing a role). Note that this condition is automatically satisfied when n is large enough if $k = O(\theta_n) = \mathrm{o}(n)$, as required by the assumptions of the lemma.

Hence for fixed $T > 0$, eventually (i.e. for n large) we have

$$\begin{aligned}
\frac{\lfloor\theta_n t\rfloor}{\sqrt{\theta_n}}\left(\frac{M^1_{\theta_n,n}(t)}{\gamma} - 1\right) \leq &\frac{1}{\sqrt{\theta_n}}\sum_{i=0}^{\lfloor\theta_n t\rfloor-1}\left(\log\frac{Y_{n-i,n}}{Y_{n-\lfloor\theta_n t\rfloor,n}} - 1\right) \\
&+\frac{1}{\rho}\sqrt{\theta_n}A_0(Y_{n-\lfloor\theta_n t\rfloor,n})\frac{1}{\theta_n}\sum_{i=0}^{\lfloor\theta_n t\rfloor-1}\left(\left(\frac{Y_{n-i,n}}{Y_{n-\lfloor\theta_n t\rfloor,n}}\right)^{\rho} - 1\right) \\
&+\epsilon\sqrt{\theta_n}|A_0(Y_{n-\lfloor\theta_n t\rfloor,n})|\frac{1}{\theta_n}\sum_{i=0}^{\lfloor\theta_n t\rfloor-1}\left(\left(\frac{Y_{n-i,n}}{Y_{n-\lfloor\theta_n t\rfloor,n}}\right)^{\rho+\epsilon}\right)
\end{aligned} \tag{A.2}$$

for every $t \in [0,T]$. Note that all the terms in the right hand side of (A.2) (which we interpret as 0 for $0 \le t < 1/\theta_n$) are in $D[0,T]$. Let us denote the second and the third terms by $J_n^{(2)}(t)$ and $J_n^{(3)}(t)$, correspondingly. We start with showing that

$$\sup_{0\le t\le T} |J_n^{(2)}(t)| \to 0 \ \text{ in probability as } n \to \infty. \tag{A.3}$$

Since $\dfrac{A_0(Y_{n-\lfloor \theta_n t\rfloor,n})}{A(Y_{n-\lfloor \theta_n t\rfloor,n})} \to 1$ a.s uniformly in $t \in [0,T]$, we may and will replace A_0 by A in this calculation. Further, as $|A|$ is eventually decreasing, for n large enough we have

$$|A(Y_{n-\lfloor \theta_n t\rfloor,n})| \le |A(Y_{n-\lfloor \theta_n T\rfloor,n})|$$

for all relevant t. Furthermore, by Corollary 2.2.2 in [de Haan and Ferreira, 2006],

$$\frac{\lfloor \theta_n T\rfloor}{n} Y_{n-\lfloor \theta_n T\rfloor} \xrightarrow{P} 1\,,$$

and, since A is regularly varying,

$$\frac{A(Y_{n-\lfloor \theta_n T\rfloor})}{A(\frac{n}{\lfloor \theta_n T\rfloor})} \xrightarrow{P} 1\,.$$

Putting everything together, we see that, in order to prove (A.3), it is enough to prove that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{0\le t\le T} \left| \sqrt{\theta_n} A(\frac{n}{\lfloor \theta_n T\rfloor}) \frac{1}{\theta_n} \sum_{i=0}^{\lfloor \theta_n t\rfloor -1} \left(\left(\frac{Y_{n-i,n}}{Y_{n-\lfloor \theta_n t\rfloor,n}} \right)^{\rho} - 1 \right) \right| > \zeta \right) = 0\,.$$

Let $\mu_\rho = 1/(1-\rho) = EY_1^\rho$. For large n we have

$$\mathbb{P}\left(\sup_{0\le t\le T} \left| \sqrt{\theta_n} A(\frac{n}{\lfloor \theta_n T\rfloor}) \frac{1}{\theta_n} \sum_{i=0}^{\lfloor \theta_n t\rfloor -1} \left(\left(\frac{Y_{n-i,n}}{Y_{n-\lfloor \theta_n t\rfloor,n}} \right)^{\rho} - 1 \right) \right| > \zeta \right)$$

$$\le \mathbb{P}\left(\sup_{0\le t\le T} \left| \sqrt{\theta_n} A(\frac{n}{\lfloor \theta_n T\rfloor}) \frac{1}{\theta_n} \sum_{i=0}^{\lfloor \theta_n t\rfloor -1} \left(\left(\frac{Y_{n-i,n}}{Y_{n-\lfloor \theta_n t\rfloor,n}} \right)^{\rho} - \mu_\rho \right) \right| > \zeta/2 \right)\,,$$

since

$$\sqrt{\theta_n}A(\frac{n}{\lfloor\theta_n T\rfloor}) \to 0$$

as $n \to \infty$, by the growth assumption on the sequence (θ_n) and the fact that $|A|$ is regularly varying with exponent ρ. The process above is a step function with jumps at multiples of $\frac{1}{\theta_n}$, hence largest value of the process is achieved at one of these steps. Therefore, the above probability does not exceed

$$\sum_{j=1}^{\lfloor T\theta_n\rfloor} \mathbb{P}\left(\left|\sqrt{\theta_n}A(\frac{n}{\lfloor\theta_n T\rfloor})\frac{1}{\theta_n}\sum_{i=0}^{j-1}\left(\left(\frac{Y_{n-i,n}}{Y_{n-j,n}}\right)^{\rho}-\mu_\rho\right)\right| > \zeta/2\right)$$

$$\leq \sum_{j=1}^{\lfloor T\theta_n\rfloor}\left(\sqrt{\theta_n}A(\frac{n}{\lfloor\theta_n T\rfloor})\right)^2\frac{4}{\zeta^2\theta_n^2}E\left[\sum_{i=0}^{j-1}\left(\left(\frac{Y_{n-i,n}}{Y_{n-j,n}}\right)^{\rho}-\mu_\rho\right)\right]^2.$$

by Chebyshev's inequality.

For each fixed j, by the Renyi representation, $\left\{\frac{Y_{n-i,n}}{Y_{n-j,n}}\right\}_i \stackrel{d}{=} \left\{Y^*_{j-i,j}\right\}_i$ where $Y^*_0, Y^*_1, \ldots, Y^*_{j-1}$ are, once again, i.i.d. Pareto(1) random variables. Therefore,

$$E\left[\sum_{i=0}^{j-1}\left(\left(\frac{Y_{n-i,n}}{Y_{n-j,n}}\right)^{\rho}-\mu_\rho\right)\right]^2 = j\mathrm{Var}(Y^{*\rho}).$$

By the growth assumption of the sequence (θ_n) we conclude that

$$\sum_{j=1}^{\lfloor T\theta_n\rfloor}\left(\sqrt{\theta_n}A(\frac{n}{\lfloor\theta_n T\rfloor})\right)^2\frac{4}{\zeta^2\theta_n^2}E\left[\sum_{i=0}^{j-1}\left(\left(\frac{Y_{n-i,n}}{Y_{n-j,n}}\right)^{\rho}-\mu_\rho\right)\right]^2$$

$$\leq \left(\sqrt{\theta_n}A(\frac{n}{\lfloor\theta_n T\rfloor})\right)^2\frac{4}{\zeta^2}\mathrm{Var}(Y^{*\rho})\frac{1}{\theta_n^2}\sum_{j=1}^{\lfloor T\theta_n\rfloor} j \to 0$$

as $n \to \infty$. This proves (A.3). In the same way we can show that

$$\sup_{\delta\leq t\leq T}|J_n^{(3)}(t)| \to 0 \text{ in probability as } n \to \infty. \tag{A.4}$$

Applying the corresponding lower bounds, we see that

$$\frac{\lfloor\theta_n t\rfloor}{\sqrt{\theta_n}}\left(\frac{M^1_{\theta_n,n}(t)}{\gamma}-1\right)-\frac{1}{\sqrt{\theta_n}}\sum_{i=0}^{\lfloor\theta_n t\rfloor-1}\left(\log\frac{Y_{n-i,n}}{Y_{n-\lfloor\theta_n t\rfloor,n}}-1\right)\to 0 \tag{A.5}$$

$t \in [\delta, \infty)$, uniformly on compact intervals, in probability. Next, we recall that $\log Y_1$ is a standard exponential random variable, so that the differences $\log Y_{n-i,n} - \log Y_{n-i-1,n}$, $i = 0, 1, \ldots, n-1$ are independent exponential random variables with the means $1/(i+1)$, $i = 0, 1, \ldots, n-1$. Therefore, denoting the ith of these exponential random variables by $E_i/(i+1)$, we see that for $k = 1, \ldots, n$,

$$\sum_{i=0}^{k-1} \log \frac{Y_{n-i,n}}{Y_{n-k,n}} = \sum_{i=0}^{k-1} E_i \,.$$

Therefore,

$$\frac{\lfloor \theta_n t \rfloor}{\sqrt{\theta_n}} \left(\frac{M^1_{\theta_n,n}(t)}{\gamma} - 1 \right) - \frac{1}{\sqrt{\theta_n}} \sum_{i=0}^{\lfloor \theta_n t \rfloor - 1} (E_i - 1) \to 0 \tag{A.6}$$

$t \in [\delta, \infty)$, uniformly on compact intervals, in probability. Squaring (A.1) and repeating the argument gives us

$$\frac{\lfloor \theta_n t \rfloor}{\sqrt{\theta_n}} \left(\frac{M^2_{\theta_n,n}(t)}{\gamma^2} - 2 \right) - \frac{1}{\sqrt{\theta_n}} \sum_{i=0}^{\lfloor \theta_n t \rfloor - 1} \big((E_i^2 - 2) \to 0, \tag{A.7}$$

$t \in [\delta, \infty)$, uniformly on compact intervals, in probability.

By Theorem 12.6.1 and Remark 12.6.2 in [Whitt, 2002], the statement of the lemma will follow once we check that

$$\frac{1}{\sqrt{\theta_n}} \begin{pmatrix} \sum_{i=0}^{\lfloor \theta_n t \rfloor - 1} (E_i - 1) \\ \sum_{i=0}^{\lfloor \theta_n t \rfloor - 1} ((E_i^2 - 2) \end{pmatrix} \Rightarrow \begin{bmatrix} W_1(t) \\ W_2(t) \end{bmatrix}$$

in $D^2[\delta, \infty)$, where $\big((W_1(t), W_2(t)),\, t \geq \delta \big)$ is a two-dimensional zero mean Brownian motion covariance matrix

$$\begin{bmatrix} 1 & 4 \\ 4 & 20 \end{bmatrix}.$$

This is, however, an immediate consequence of the multivariate version of Donsker's theorem; see e.g. Theorem 4.3.5 in [Whitt, 2002]. □

A.2 Proof of Theorem 1

Proof. We start by showing that for any $\delta > 0$,

$$\sqrt{t}Q_n^*(t) \Rightarrow \frac{1}{2}(W_2(t) - 2W_1(t)) \tag{A.8}$$

in $D[\delta, \infty)$, where $\big((W_1(t), W_2(t)),\, t \geq 0\big)$ is the two-dimensional Brownian motion of Lemma 2.2.1, and

$$Q_n^*(t) = \begin{cases} Q_{\lfloor \theta_n t \rfloor, n} & \text{if } \ \delta \leq t \leq \dfrac{n}{\theta_n}, \\ Q_{n,n} & \text{if } \ t > \dfrac{n}{\theta_n}\,. \end{cases}$$

By Theorem 16.7 in [Billingsley, 1999], we have to prove convergence in $D[\delta, T]$ for each $\delta < T < \infty$. Straightforward algebra shows that

$$\sqrt{t}Q_n^*(t) = \frac{\sqrt{\theta_n}t}{2} \frac{\big(M^2_{\theta_n,n}(t)/\gamma^2 - 2\big) - 4\big(M^1_{\theta_n,n}(t)/\gamma - 1\big)}{\big(M^1_{\theta_n,n}(t)\big)^2/\gamma^2}$$

$$-\sqrt{\theta_n}t \frac{\big(M^1_{\theta_n,n}(t)/\gamma - 1\big)^2}{\big(M^1_{\theta_n,n}(t)\big)^2/\gamma^2} := V_n^{(1)}(t) - V_n^{(2)}(t)\,,$$

$\delta \leq t \leq T$, while for $0 \leq t < 1/n$ we define both $V_n^{(1)}(t) = 0$ and $V_n^{(2)}(t) = 0$. Call

$$D_n^{(1)}(t) = \frac{\sqrt{\theta_n}t}{2}\Big[\big(M^2_{\theta_n,n}(t)/\gamma^2 - 2\big) - 4\big(M^1_{\theta_n,n}(t)/\gamma - 1\big)\Big],\ \delta \leq t \leq T.$$

Since the limiting process in Lemma 2.2.1 is continuous, the weak convergence holds also in the uniform topology (on $[\delta, T]$), and addition is continuous in this topology. By Lemma 2.2.1 and continuous mapping theorem we conclude that

$$D_n^{(1)}(t) \Rightarrow \frac{1}{2}(W_2(t) - 4W_1(t))$$

in $D[\delta, T]$. Furthermore, by Lemma 2.2.1,

$$M^1_{\theta_n,n}(t)/\gamma \to 1$$

uniformly on $[\delta, T]$ in probability. By Theorem 3.1 in [Billingsley, 1999] we conclude that

$$V_n^{(1)}(t) \Rightarrow \frac{1}{2}(W_2(t) - 4W_1(t))$$

in $D[\delta, T]$ as well. Similarly,

$$V_n^{(2)}(t) \to 0$$

uniformly on $[\delta, T]$ in probability, so that we may apply Theorem 3.1 in [Billingsley, 1999] once again and obtain (A.8).

Fix $x > 0$, and write

$$P(N_n \le \theta_n x) = P\left(|Q_{k,n}| \ge \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } 1 \le k \le \theta_n x \right).$$

Therefore, for $0 < \delta < x$ we have

$$P\left(|Q_{k,n}| \ge \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } \theta_n\delta \le k \le \theta_n x \right) \le P(N_n \le \theta_n x)$$

$$\le P\left(|Q_{k,n}| \ge \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } \theta_n\delta \le k \le \theta_n x \right)$$

$$+P\left(|Q_{k,n}| \ge \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } 1 \le k \le \theta_n\delta \right).$$

We will show that for any $0 < \delta < x$

$$\lim_{n\to\infty} P\left(|Q_{k,n}| \ge \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } \theta_n\delta \le k \le \theta_n x \right) = P\big(\sup_{\delta\le t\le x} |B(t)| \ge \omega \big), \quad \text{(A.9)}$$

where $\big(B(t),\, t \ge 0\big)$ is a standard Brownian motion, while

$$\lim_{\delta\to 0} \limsup_{n\to\infty} P\left(|Q_{k,n}| \ge \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } 1 \le k \le \theta_n\delta \right) = 0\,. \quad \text{(A.10)}$$

It will follow from the above relations that

$$P(N_n \le \theta_n x) \to P\big(\sup_{0\le t\le x} |B(t)| \ge \omega \big) = P(\tau_\omega \le x),$$

which is what we need for the statement of the theorem.

Observe that for any $\delta > 0$,

$$P\left(|Q_{k,n}| \geq \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } \theta_n\delta \leq k \leq \theta_n x\right)$$

$$= P\left(|\sqrt{t}Q_n^*(t)| \geq \sqrt{t}\omega\sqrt{\frac{\theta_n}{[\theta_n t]}} \text{ for some } \delta \leq t \leq x\right)$$

$$= P\left(|\sqrt{t}Q_n^*(t)| \geq \omega(1+o(1)) \text{ for some } \delta \leq t \leq x\right)$$

(with the same $o(1)$ for all relevant t). Now (A.9) follows from (A.8) and the continuity of the supremum distribution of the Brownian motion.

In order to show (A.10), we start with showing that, for any $\delta > 0$, both

$$\inf_{1\leq k\leq \theta_n\delta} M_{k,n}^1 \text{ is stochastically bounded away from 0, and} \tag{A.11}$$

$$\sup_{1\leq k\leq \theta_n\delta} M_{k,n}^1 \text{ is stochastically bounded away from infinity.}$$

To see this, we recall from the decomposition in (A.2) and the subsequent argument in the proof of Lemma 2.2.1 that

$$\left(M_{k,n}^1,\, 1 \leq k \leq \theta_n\delta\right) = \left(M_{k,n}^{1,Y}\gamma + W_{k,n},\, 1 \leq k \leq \theta_n\delta\right),$$

where $(M_{k,n}^{1,Y})$ is constructed using the standard Pareto random variables, while

$$\sup_{1\leq k\leq \theta_n\delta} |W_{k,n}| \to 0 \text{ in probability.}$$

Therefore, (A.11) will follow once we check it for the standard Pareto random variables. However, we have seen that, in the latter case,

$$\left(M_{k,n}^{1,Y},\, 1 \leq k \leq \theta_n\delta\right) \stackrel{d}{=} \left(\frac{1}{k}\sum_{i=1}^{k} E_i,\, 1 \leq k \leq \theta_n\delta\right),$$

where $E_1, E_2, \ldots$ are i.i.d. standard exponential random variables, and, hence, (A.11) follows from the law of large numbers.

We continue by writing

$$P\left(|Q_{k,n}| \geq \omega\sqrt{\frac{\theta_n}{k}} \text{ for some } 1 \leq k \leq \theta_n\delta\right)$$

$$= P\left(k\left|\frac{M^2_{k,n} - 2(M^1_{k,n})^2}{2(M^1_{k,n})^2}\right| > \omega\sqrt{\theta_n} \text{ for some } 1 \leq k \leq \theta_n\delta\right)$$

and, using (A.11), we see that, in order to prove (A.10), it is enough to prove that

$$\lim_{\delta\to 0}\limsup_{n\to\infty} P\left(|R_{k,n}| \geq \omega\sqrt{\theta_n} \text{ for some } 1 \leq k \leq \theta_n\delta\right) = 0\,, \tag{A.12}$$

where

$$R_{k,n} = k\frac{M^2_{k,n} - 2(M^1_{k,n})^2}{2(M^1_{k,n})^2},\ k = 1, 2, \ldots, n\,.$$

A straightforward algebra allows us to write the above probability as

$$P\left(\frac{k}{2}\left|\Big(\frac{M^2_{k,n}}{\gamma^2} - 2\Big) - 4\Big(\frac{M^1_{k,n}}{\gamma} - 1\Big)\right.\right.$$

$$\left.\left. -2\Big(\frac{M^1_{k,n}}{\gamma} - 1\Big)^2\right| > \omega\sqrt{\theta_n} \text{ for some } 1 \leq k \leq \theta_n\delta\right)\,.$$

By (A.11), we can write now

$$P\left(|R_{k,n}| \geq \omega\sqrt{\theta_n} \text{ for some } 1 \leq k \leq \theta_n\delta\right)$$

$$\leq P\Bigg(\frac{k}{2}\left(\left|\frac{M^2_{k,n}}{\gamma^2} - 2\right| + 4\left|\frac{M^1_{k,n}}{\gamma} - 1\right|\right)(1 + K_n) \geq \omega\sqrt{\theta_n}$$

$$\text{for some } 1 \leq k \leq \theta_n\delta\Bigg)\,,$$

where (K_n) is a tight family of nonnegative random variables. Since by Lemma 2.2.1,

$$P\Bigg(\frac{k}{2}\left(\left|\frac{M^2_{k,n}}{\gamma^2} - 2\right| + 4\left|\frac{M^1_{k,n}}{\gamma} - 1\right|\right) \geq \omega\sqrt{\theta_n} \text{ for some } 1 \leq k \leq \theta_n\delta\Bigg)$$

$$= P\Bigg(\sup_{0\leq t\leq\delta}\sqrt{\theta_n}t\Big(\left|\frac{M^2_{[\theta_n t],n}}{\gamma^2} - 2\right| + 4\left|\frac{M^1_{[\theta_n t],n}}{\gamma} - 1\right|\Big) > \omega\Bigg)$$

$$\to P(\sup_{0\leq t\leq\delta}|B(t)| > \omega)\,,$$

where (B) is some Brownian motion, the claim (A.10) follows. This completes the proof of the theorem. □

A.3 Proof of Theorem 2

Proof. The idea is to use a random stopping technique in a weak convergence context. The formulation we will use is the one given in Theorem 2.2.1 in [Silvestrov, 2004]. If, for each n, $(X_n(t),\, t \geq 0)$ is a càdlàg process, and τ_n is a nonnegative random variable, such that for all $0 \leq a < b < \infty$,

$$\begin{aligned} (\tau_n,\, \sup_{t\in[a,b)} X_n(t)) &\Rightarrow (\tau_0,\, \sup_{t\in[a,b)} X_0(t)) \\ (\tau_n,\, \inf_{t\in[a,b)} X_n(t)) &\Rightarrow (\tau_0,\, \inf_{t\in[a,b)} X_0(t)) \end{aligned} \tag{A.13}$$

for some continuous process $(X_0(t),\, t \geq 0)$ and a nonnegative random variable τ_0, then $X_n(\tau_n) \Rightarrow X_0(\tau_0)$.

It is, clearly, enough to prove the weak convergence (2.4), as the consistency of the estimator would then follow automatically. Note that (2.4) is equivalent to

$$\sqrt{\theta_n}\frac{1}{N_n}\sum_{i=0}^{N_n-1}\Big(\frac{1}{\gamma}\log\frac{X_{n-i,n}}{X_{n-N_n,n}} - 1\Big) \Rightarrow \frac{G}{(\tau_\omega)^{1/2}}\,. \tag{A.14}$$

We will prove that, for any $\delta > 0$,

$$\sqrt{\theta_n}\frac{1}{N_n \vee \theta_n\delta}\sum_{i=0}^{[N_n\vee\theta_n\delta]-1}\Big(\frac{1}{\gamma}\log\frac{X_{n-i,n}}{X_{n-[N_n\vee\theta_n\delta],n}} - 1\Big) \Rightarrow \frac{G}{(\tau_\omega \vee \delta)^{1/2}}\,. \tag{A.15}$$

The claim (A.14) would then follows from (A.10).

Note the that expression in the left hand side of (A.15) results from a substitution of the random time

$$\tau_n = \max\left(\delta, \frac{N_n}{\theta_n}\right)$$

into the càdlàg process

$$V_n(t) = \sqrt{\theta_n}\frac{1}{\theta_n t}\sum_{i=0}^{[\theta_n t]-1}\Big(\frac{1}{\gamma}\log\frac{X_{n-i,n}}{X_{n-[\theta_n t],n}} - 1\Big),\; t \geq \delta\,.$$

According to (A.13) and to self-similarity of the Brownian motion, it is enough to check that for all $\delta \le a < b < \infty$,

$$\left(\max\left(\delta, \frac{N_n}{\theta_n}\right), \sup_{t\in[a,b)} V_n(t)\right) \Rightarrow \left(\tau_\omega \vee \delta, \sup_{t\in[a,b)} \frac{B(t)}{t}\right), \tag{A.16}$$

where τ_ω is, as before, the first hitting time of a standard Brownian motion, independent of another standard Brownian motion B. We also need to prove an analogous statement with suprema replaced by infima but, since the two statements can be proved in the same way, we concentrate on suprema only.

In order to prove (A.16), it is enough to show that for any $x \ge \delta$ and $y \ge 0$,

$$P\Big(N_n \le \theta_n x, \sup_{t\in[a,b)} V_n(t) \le y\Big) \to P\Big(\tau_\omega \le x, \sup_{t\in[a,b)} \frac{B(t)}{t} \le y\Big).$$

We have seen in an analogous situation in the proof of Theorem 1 that this statement will follow once we check that for any $0 < \delta' < x$,

$$P\Big(\theta_n\delta' \le N_n \le \theta_n x, \sup_{t\in[a,b)} V_n(t) \le y\Big)$$

$$\to P\Big(\sup_{\delta'\le t\le x} |B_1(t)| \ge \omega, \sup_{t\in[a,b)} \frac{B(t)}{t} \le y\Big),$$

where B and B_1 are independent standard Brownian motions, which, in turn, will be implied by the statement

$$P\Big(\sup_{\delta'\le t\le x} |\sqrt{t}\, Q_n^*(t)| \ge w, \sup_{t\in[a,b)} V_n(t) \le y\Big) \tag{A.17}$$

$$\to P\Big(\sup_{\delta'\le t\le x} |B_1(t)| \ge \omega, \sup_{t\in[a,b)} \frac{B(t)}{t} \le y\Big).$$

To this end note that

$$V_n(t) = \frac{1}{t}\left[\sqrt{\theta_n}\, t\left(\frac{M^1_{\theta_n,n}(t)}{\gamma} - 1\right)\right],$$

and the map $\big(f(t),\, t \geq \delta\big) \to \big(f(t)/t,\, t \geq \delta\big)$ is continuous on $D[\delta, \infty)$. Therefore, the argument leading to (A.8) applies, and it gives us the joint convergence

$$\begin{bmatrix} \sqrt{t}\, Q_n^*(t),\, t \geq \delta' \\ V_n(t),\, t \geq \delta' \end{bmatrix} \Rightarrow \begin{bmatrix} \frac{1}{2}\big(W_2(t) - 4W_1(t)\big),\, t \geq \delta' \\ \frac{W_1(t)}{t},\, t \geq \delta' \end{bmatrix},$$

where W_1 and W_2 are as in Lemma 2.2.1. It is a simple matter to compute the correlations and check that W_1 and $(W_2 - 4W_1)/2$ are independent standard Brownian motions. Therefore, (A.17) follows, and the proof of the theorem is complete. □

A.4 Proof of Theorem 4

Proof. It is enough to show that for each $j = 1, \ldots, d$ and $t > 0$,

$$\frac{X^{(j)}_{\lceil N_n t \rceil}}{b^{(j)}\big(\frac{n}{N_n}\big)} \to t^{-1/\alpha_j} \text{ in probability,} \tag{A.18}$$

because then the argument on pp. 311-312 in [Resnick, 2007] can be used to establish the claim of the theorem. We start with checking that for every $j, m = 1, \ldots, d$

$$\frac{X^{(j)}_{\lceil N_n^{(m)} t \rceil}}{b^{(j)}\big(\frac{n}{N_n^{(m)}}\big)} \to t^{-1/\alpha_j} \text{ in probability.} \tag{A.19}$$

Fix j, m and let $0 < a < A < \infty$. We verify first that

$$\left(\frac{X^{(j)}_{\lceil \lambda_n^{(m)} u t \rceil}}{b^{(j)}\big(\frac{n}{\lambda_n^{(m)} u}\big)},\, a \leq u \leq A \right) \to \big(t^{-1/\alpha_j} e(u),\, a \leq u \leq A\big) \tag{A.20}$$

in probability in the Skorohod J_1 topology on $D[a, A]$. Here e is a function on $[a, A]$ equal identically to 1.

Indeed, let $0 < \varepsilon < t^{-1/\alpha_j}$, and choose K so large that

$$\frac{u_i}{u_{i-1}} < (1 + \varepsilon t^{1/\alpha_j})^{\alpha_j} \quad \text{for each } i = 1, \ldots, k,$$

with $u_i = a + (A-a)i/K,\ i = 0, 1, \ldots, K$. Then

$$P\left(\sup_{a\le u\le A}\left|\frac{X^{(j)}_{\lceil\lambda_n^{(m)}ut\rceil}}{b^{(j)}(\frac{n}{\lambda_n^{(m)}u})} - t^{-1/\alpha_j}\right| > \varepsilon\right)$$

$$\le P\Big(X^{(j)}_{\lceil\lambda_n^{(m)}ut\rceil} > \big(t^{-1/\alpha_j}+\varepsilon\big)b^{(j)}\big(\frac{n}{\lambda_n^{(m)}u}\big) \ \text{ for some } a\le u\le A\Big)$$

$$+P\Big(X^{(j)}_{\lceil\lambda_n^{(m)}ut\rceil} < \big(t^{-1/\alpha_j}-\varepsilon\big)b^{(j)}\big(\frac{n}{\lambda_n^{(m)}u}\big) \ \text{ for some } a\le u\le A\Big)$$

$$:= a_n + b_n\,.$$

By monotonicity,

$$a_n \le \sum_{i=1}^{K} P\Big(X^{(j)}_{\lceil\lambda_n^{(m)}u_{i-1}t\rceil} > \big(t^{-1/\alpha_j}+\varepsilon\big)b^{(j)}\big(\frac{n}{\lambda_n^{(m)}u_i}\big)\Big)\,,$$

and the ith term in the finite sum can be written as

$$P\left(\frac{X^{(j)}_{\lceil\lambda_n^{(m)}u_{i-1}t\rceil}}{b^{(j)}(\frac{n}{\lambda_n^{(m)}u_{i-1}})} > \big(t^{-1/\alpha_j}+\varepsilon\big)\frac{b^{(j)}(\frac{n}{\lambda_n^{(m)}u_i})}{b^{(j)}(\frac{n}{\lambda_n^{(m)}u_{i-1}})}\right) \to 0$$

as $n\to\infty$, because by (4.18) in [Resnick, 2007], the random variable in the left hand side of the inequality converges in probability to t^{-1/α_j}, while the expression in the right hand side of the inequality converges, by the regular variation, to

$$\big(t^{-1/\alpha_j}+\varepsilon\big)\left(\frac{u_{i-1}}{u_i}\right)^{1/\alpha_j} > t^{-1/\alpha_j}$$

by the choice of K. Hence $a_n\to 0$ and, similarly, $b_n\to 0$ as $n\to\infty$, which establishes (A.20).

Let $\varphi:\,(0,\infty)\to[a,A]$ be defined by

$$\varphi(x) = \begin{cases} x, & a\le x\le A \\ a, & 0<x<a \\ A, & x>A. \end{cases}$$

Since φ is a continuous function, and $N_n^{(m)}/\lambda_n^{(m)} \Rightarrow \tau_{\omega_m}$ by Theorem 1 (recall that τ_{ω_m} is the first time a standard Brownian motion hits $\pm\omega_m$), we conclude that

$\varphi\big(N_n^{(m)}/\lambda_n^{(m)}\big) \Rightarrow \varphi\big(\tau_{\omega_m}\big)$. Since the convergence in (A.20) is to a deterministic limit, it follows by Theorem 4.4 in [Billingsley, 1968] that

$$\left[\left(\frac{X^{(j)}_{\lceil \lambda_n^{(m)} ut\rceil}}{b^{(j)}\big(\frac{n}{\lambda_n^{(m)}u}\big)},\, a \le u \le A\right),\, \varphi\big(N_n^{(m)}/\lambda_n^{(m)}\big)\right]$$

$$\Rightarrow \Big[\big(t^{-1/\alpha_j} e(u),\, a \le u \le A\big),\, \varphi\big(\tau_{\omega_m}\big)\Big]$$

weakly in $D[a,A] \times [a,A]$. Since the map $(x,t) \to x(t)$ on $D[a,A] \times [a,A]$ is continuous at every point of its domain with a continuous first coordinate, we conclude that

$$\frac{X^{(j)}_{\lceil \lambda_n^{(m)} \varphi\left(N_n^{(m)}/\lambda_n^{(m)}\right)\varphi\left(\tau_{\omega_m}\right) t\rceil}}{b^{(j)}\big(\frac{n}{\lambda_n^{(m)}\varphi\left(N_n^{(m)}/\lambda_n^{(m)}\right)}\big)} \to t^{-1/\alpha_j}$$

in probability. Since

$$P\left(\frac{X^{(j)}_{\lceil N_n^{(m)} t\rceil}}{b^{(j)}\big(\frac{n}{N_n^{(m)}}\big)} \neq \frac{X^{(j)}_{\lceil \lambda_n^{(m)} \varphi\left(N_n^{(m)}/\lambda_n^{(m)}\right)\varphi\left(\tau_{\omega_m}\right) t\rceil}}{b^{(j)}\big(\frac{n}{\lambda_n^{(m)}\varphi\left(N_n^{(m)}/\lambda_n^{(m)}\right)}\big)}\right) \le P\left(\frac{N_n^{(m)}}{\lambda_n^{(m)}} \notin [a,A]\right)$$

$$\to P\big(\tau_{\omega_m} \notin [a,A]\big)\,,$$

which can be made arbitrarily small by taking a small and A large, (A.19) follows.

The connection, through subsequences, between convergence in probability and a.s. convergence shows that (A.19) implies (A.18), so that the proof of the theorem is complete. □

BIBLIOGRAPHY

Aas, K. and Haff, I. H. (2006). The generalized hyperbolic skew students t-distribution. *Journal of Financial Econometrics*, 4(2):275–309.

Adler, R., Feldman, R., and Taqqu, M. (1998). *A practical guide to heavy tails: statistical techniques and applications*. Birkhäuser Boston.

Adrian, T. and Brunnermeier, M. (2011). Covar. Technical report, National Bureau of Economic Research.

Arzac, E. R. and Bawa, V. S. (1977). Portfolio choice and equilibrium in capital markets with safety-first investors. *Journal of Financial Economics*, 4(3):277–288.

Baillie, R. T. and Bollerslev, T. (1989). The message in daily exchange rates: a conditional-variance tale. *Journal of Business & Economic Statistics*, 7(3):297–305.

Baillie, R. T. and DeGennaro, R. P. (1990). Stock returns and volatility. *Journal of financial and quantitative analysis*, 25(2):203–214.

Barro, R. J. (2006). Rare disasters and asset markets in the twentieth century. *The Quarterly Journal of Economics*, 121(3):823–866.

Bauwens, L. and Laurent, S. (2005). A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *Journal of Business & Economic Statistics*, 23(3).

Beine, M., Laurent, S., and Lecourt, C. (2002). Accounting for conditional leptokurtosis and closing days effects in figarch models of daily exchange rates. *Applied Financial Economics*, 12(8):589–600.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes*. Wiley Online Library.

Bellman, R., Bellman, R. E., and Bellman, R. E. (1966). *Adaptive control processes: A guided tour*. Princeton University Press.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, New York, second edition.

Bocker, K. and Kluppelberg, C. (2005). Operational VaR: a closed-form approximation. *RISK-LONDON-RISK MAGAZINE LIMITED-*, 18(12):90.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, pages 542–547.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2000). Copulas for finance-a reading guide and some applications. *Available at SSRN 1032533*.

Chen, S. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2):131–145.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. Wiley.

Dahiya, R. and Gurland, J. (1972). Goodness of fit tests for the gamma and exponential distributions. *Technometrics*, 14:791–801.

Danielsson, J., De Haan, L., Peng, L., and De Vries, C. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. *Journal of Multivariate analysis*, 76(2):226–248.

Danielsson, J. and de Vries, C. G. (1997). Tail index and quantile estimation with very high frequency data. *Journal of Empirical Finance*, 4(2):241–257.

Danielsson, J. and De Vries, C. G. (2000). Value-at-risk and extreme returns. *Annales d'Economie et de Statistique*, pages 239–270.

Danielsson, J. and Vries, C. d. (1996). Beyond the sample: Extreme quantile and probability estimation. Technical report, Tinbergen Institute.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer, New York.

de Haan, L. and Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica*, 52:60–70.

Deheuvels, P., Häusler, E., and Mason, D. (1988). Almost sure convergence of the hill estimator. *Mathematical Proceedings of Cambridge Philosophical Society*, 104:371–381.

Dekkers, A. and de Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *Annals of Statistics*, 17:1795–1832.

Dekkers, A., Einmahl, J., and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, 17:1833–1855.

Drees, H., de Haan, L., and Resnick, S. (2000). How to make a Hill plot. *Annals of Statistics*, 28(1):254–274.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75:149–172.

Duong, T. and Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506.

Einmahl, J., de Haan, L., and Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *Annals of Statistics*, 9:1401–1423.

Einmahl, J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989.

Einmahl, J. H., De Haan, L., and Huang, X. (1993). Estimating a multidimensional extreme-value distribution. *Journal of multivariate analysis*, 47:35–35.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, pages 176–223.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, volume 2. Wiley, New York, 2nd edition.

Geluk, J. and de Haan, L. (1987). *Regular Variation, Extensions and Tauberian Theorems*. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam. Volume 40 of *CWI Tract*.

Girardi, G. and Ergun, A. T. (2012). Systemic risk measurement: Multivariate garch estimation of covar. *Available at SSRN 1783958*.

Glasserman, P. (2003). *Monte Carlo methods in financial engineering*, volume 53. Springer.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate analysis*, 32(2):177–203.

Hall, P., Marron, J., and Park, B. U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, 92(1):1–20.

Hall, P. and Welsh, A. (1985). Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13:331–341.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, pages 705–730.

Hauksson, H., Dacorogna, M., Domenig, T., Müller, U., and Samorodnitsky, G. (2001). Multivariate extremes, aggregation and risk estimation. *Quantitative Finance*, 1:79–95.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349.

Hsing, T. (1991). On tail index estimation using dependent data. *The Annals of Statistics*, 19:1547–1569.

Hyung, N. and de Vries, C. G. (2005). Portfolio diversification effects of downside risk. *Journal of Financial Econometrics*, 3(1):107–125.

Hyung, N. and de Vries, C. G. (2007). Portfolio selection with heavy tails. *Journal of Empirical Finance*, 14(3):383–400.

Ibragimov, R. and Walden, J. (2007). The limits of diversification when losses may be large. *Journal of Banking & Finance*, 31(8):2551–2569.

Jansen, D. W. and De Vries, C. G. (1991). On the frequency of large stock returns: Putting booms and busts into perspective. *The review of economics and statistics*, pages 18–24.

Jansen, D. W., Koedijk, K. G., and De Vries, C. G. (2000). Portfolio selection with limited downside risk. *Journal of Empirical Finance*, 7(3):247–269.

Kearns, P. and Pagan, A. (1997). Estimating the density tail index for financial time series. *Review of Economics and Statistics*, 79(2):171–175.

Khindanova, I., Rachev, S., and Schwartz, E. (2001). Stable modeling of value at risk. *Mathematical and Computer Modelling*, 34(9):1223–1259.

Klyman, J. (2011). *Systemic Risk Measures: DistVar and Other "Too Big To fail" Risk Measures*. PhD thesis, Princeton University.

Ledford, A. and Tawn, J. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187.

Ledford, A. and Tawn, J. (1997). Modelling dependence within joint tail regions. *Journal of Royal Statistical Society, Series B*, 59:475–499.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The review of economics and statistics*, 47(1):13–37.

Longin, F. and Solnik, B. (2001). Extreme correlation of international equity markets. *The Journal of Finance*, 56(2):649–676.

Loretan, M. and Phillips, P. (1994). Testing the covariance stationarity of heavy-tailed time series. *J. Empirical Finance*, 1:211–248.

Mainik, G. and Schaanning, E. (2012). On dependence consistency of covar and some other systemic risk measures. *arXiv preprint arXiv:1207.3464*.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 26:394–419.

Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1):77–91.

Markowitz, H. (1959). Portfolio selection: efficient diversification of investments. Cowles Foundation Monograph No. 16.

Mason, D. (1982). Laws of large numbers for sums of extreme values. *Annals of Probability*, 10:756–764.

McCulloch, J. H. (1996). Financial applications of stable distributions. *Handbook of statistics*, 14:393–425.

McNeil, A. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27:117–137.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, 7(3):271–300.

Mikosch, T. (2006). Copulas: Tales and facts. *Extremes*, 9(1):3–20.

Mittnik, S., Rachev, S., and Schwartz, E. (2002). Value-at-risk and asset allocation with stable return distributions. *Allgemeines Statistisches Archiv*, 86(1):53–67.

Mittnik, S., Rachev, S. T., and Paolella, M. S. (1998). Stable paretian modeling in finance: Some empirical and theoretical aspects. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications. Boston: Birkhaüser*, pages 79–110.

Moscadelli, M. (2004). The modelling of operational risk: experience with the analysis of the data collected by the basel committee. *Available at SSRN 557214*.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica: Journal of the Econometric Society*, pages 768–783.

Neves, C. and Fraga Alves, M. (2007). Semi-parametric approach to the hasofer–wang and greenwood statistics in extremes. *Test*, 16(2):297–313.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Stat.*, 3:119–131.

Rachev, S., editor (2003). *Handbook of Heavy Tailed Distributions in Finance*. Elsevier, Amsterdam.

Rachev, S. and Mittnik, S. (2000). *Stable Paretian Models in Finance*. Wiley, Chechester, New York.

Rachev, S. T., Menn, C., Fabozzi, F. J., et al. (2005). *Fat-tailed and skewed asset return distributions: implications for risk management, portfolio selection, and option pricing*, volume 139. Wiley.

Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York.

Resnick, S. (1997). Discussion of the danish data on large fire insurance losses. *ASTIN Bulletin*, 27:139–151.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York.

Resnick, S. and Stărică, C. (1998). Tail index estimation for dependent data. *Annals of Applied Probability*, 8:1156–1183.

Resnick, S. and Stărică, C. (1995). Consistency of hill estimator for dependent data. *Journal of Applied Probability*, 32:139–167.

Resnick, S. and Stărică, C. (1997). Smoothing the Hill estimator. *Advances in Applied Probability*, 20:271–293.

Rootzén, H. and Tajvidi, N. (1997). Extreme value statistics and wind storm losses: a case study. *Scandinavian Actuarial Journal*, 1997(1):70–94.

Roy, A. D. (1952). Safety first and the holding of assets. *Econometrica: Journal of the Econometric Society*, pages 431–449.

Samorodnitsky, G. and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes*. Chapman and Hall, New York.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk*. *The journal of finance*, 19(3):425–442.

Silvestrov, D. (2004). *Limit theorems for randomly stopped stochastic processes*. Springer, London.

Stephenson, A. and Tawn, J. (2005). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika*, 92(1):213–227.

Stephenson, A. G. (2009). High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51(1):77–88.

Stoyanov, S., Samorodnitsky, G., Rachev, S., and Ortobelli Lozza, S. (2006). Computing the portfolio conditional value-at-risk in the alpha-stable case. *Probability and Mathematical Statistics*, 26(1):1–22.

Stărică, C. (1999). Multivariate extremes for models with constant conditional correlations. *Journal of Empirical Finance*, 6:515–553.

Susmel, R. (2001). Extreme observations and diversification in latin american emerging equity markets. *Journal of International Money and Finance*, 20(7):971–986.

Tawn, J. (1988). Bivariate extreme value theory: Models and estimation. *Biometrica*, 75:397–415.

Theodossiou, P. (1998). Financial data and the skewed generalized t distribution. *Management Science*, 44(12-Part-1):1650–1661.

Trivedi, P. K. and Zimmer, D. M. (2007). *Copula modeling: an introduction for practitioners*, volume 1. Now Publishers.

Wand, M. and Jones, C. (1994a). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116.

Wand, M. P. and Jones, M. C. (1994b). *Kernel smoothing*, volume 60. Chapman & Hall/CRC.

Wang, J. (1995). Selection of the k largest order statistics for the domain of attraction of the gumbel distribution. *Journal of the American Statistical Association*, 90(431):1055–1061.

Whitehouse, M. (2005). How a formula ignited market that burned some big investors. September 12.

Whitt, W. (2002). *Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and Their Applications to Queues*. Springer, New York.

Zhou, C. (2010). Are banks too big to fail? measuring systemic importance of financial institutions. *International Journal of Central Banking*, 6(34):205–250.